THE UNIVERSITY OF CHICAGO

INDEXING A MENTAL CHARACTERISTIC

A DISSERTATION
SUBMITTED TO THE FACULTY
OF THE GRADUATE SCHOOL OF ARTS AND LITERATURE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATION

BY
KARL JOHN HOLZINGER

CHICAGO, ILLINOIS
SEPTEMBER, 1922

TABLE OF CONTENTS

Section 1. Introduction

A. Mental and Physical Measurement

Most of the difficulties in mental measurement are due to the fact that the method is necessarily indirect. This implies vagueness not only in the thing measured, but also in the precision of the result. Many physical measurements on the other hand are direct, making possible a clearer definition of the traits measured and greater accuracy of their determination. The case of a boy to be weighed and also tested for intelligence will bring out this contrast.

If the boy is placed on the physical scale and weighed, the measurement is recorded directly in objectively defined units, accuracy of determination depending upon the instrument and the individual making the measurement. The reaction of the boy himself at the time of the weighing will have no effect on the result. In case a mental measurement of intelligence is required, it is necessary to submit to the boy a series of questions to which he must respond before a score for the mental trait may be obtained. This very indirectness of approach leads to ambiguity regarding the trait measured, and to uncertainty in the precision of the score. Thus in passing from physical to mental measurement the role of the boy has shifted from a passive to an active one, thereby complicating the whole procedure.

It may be pointed out briefly that this procedure in the case of mental measurements has very closely paralled

the physical method. Units have been defined and scales constructed. These units for the most part are functions of group variability on particular types of material, and consequently lose in accuracy and significance when applied in individual measurement. The parallelism in method has even been carried so far as to attempt to determine "zero points" for certain mental abilities so that " just not any" amounts of the traits could be used as reference points. In the case of the measurement of heat, the zero on the Fahrenheit thermometer does not mean "just no heat," but nevertheless serves as a convenient reference point. Similarly a good many of the former nebulous zero points in mental measurement have disappeared, or have moved up to the median where they belong.

b. A General Statistical Theory

The difficulties briefly sketched in the above paragraphs suggest the desirability of some more general method of approach to the problem of mental measurement, and indeed such a method has been implicitly used in some of the later work on test reliability and validity. In order to fix ideas, therefore, certain terms will now be defined so that their meaning will be clear throughout the subsequent discussion.

The term characteristics will be used to denote the physical, mental, or social traits which individuals of a group have in common. The group may consist of a number of persons or things each of which must possess the characteristic in question before any statistical study is possible. A man of

certain height, intelligence, and wealth furnishes an example of an individual with the three types of characteristics commonly studied.

The phase of a characteristic may be briefly described as the status of an individual with respect to the characteristic. This conception is an important one because it is introduced for the purpose of distinguishing a particular thing from the number that may be attached to it. Phases may be numerically or verbally expressed. Thus in describing the characteristics height and political affiliation of a certain man, the number 68 may be attached to the phase in height and the word "Republican" to the phase in political affiliation. It is conceivable that a numerical scheme for the latter might be worked out, but the phases of political affiliation per se would of course remain unchanged.

In order that a characteristic be numerically indexed, it is desirable that its phases be arranged in some order e.g. like the points on a line. If the linear arrangement be, made, the trait may be termed a linear ordinal characteristic. For characteristics such as height, an infinite number of phases are assumed, indexed by the real number system (dense set). In the case of such characteristics as size of school class, however, the number of phases is finite, and the indexing is accomplished by assigning only integers. The distinction is essentially that between continuous and discrete series, the continuity and discreetness appearing in phase.

Finally an index variable will be described as a quan-

tity whose values are in one-to-one correspondence with the phases of the characteristics indexed. Before proceeding farther, the meaning of these various terms will be illustrated by means of an artificial example.

c. A Suppostitious Example

Consider a set of 90 cubes of homogeneous material. The problem is to describe these cubes by ordinary statistical procedure. Assuming that the size of the cube is the characteristic to be studied, some mode of indexing or index variable must be adopted. Taking the edge as a first choice, the distribution may be given as follows:

edge	frequency
1	10
2	20
3	30
4	20
5	10

The mean edge is clearly 3, with corresponding face area 9 and volume 27 i.e. $a = e^2$, $v = e^3$.

A second mode of indexing by the area of a face gives the distribution:

area	frequency	fa
1	10	10
4	20	80
9	30	270
16	20	320
25	10	250
		950

The mean face area is now $10\frac{1}{3}$ with a corresponding edge and volume approximately 3.2 and 33.3 respectively.

Again indexing by volume gives:

volume	frequency	fv
1	10	10
8	20	160
27	30	810
64	20	1280
125	10	1250
		3510

The mean volume is 39, with the corresponding edge and area approximately 3.4 and 11.5. These results may be set forth in summary form as follows:

TABLE 1.- MEANS AND CORRESPONDING VALUES FOR CUBES INDEXED BY EDGE, AREA, AND VOLUME

Mode of Indexing or Index Variable	Mean and Corresponding Indexes		
	Edge	Area	Volume
Edge	3	9	27
Area	3.2	10.3	33.2
Volume	3.4	11.5	39

Inspection of these figures reveals a complete inconsistency under the three modes of indexing. Moreover the three distributions are quite different. The frequency polygon according to edge is symmetrical, while the distributions for area and volume are skewed toward the smaller values of the variable. It is also clear that the cubes remained at the same phases in the characteristic size but that various modes of indexing gave inconsistent results. The above example then illustrates the fact that although a set of things be unaltered in phase, the form of the distribution and the statistical constants depend upon the particular index variable employed.

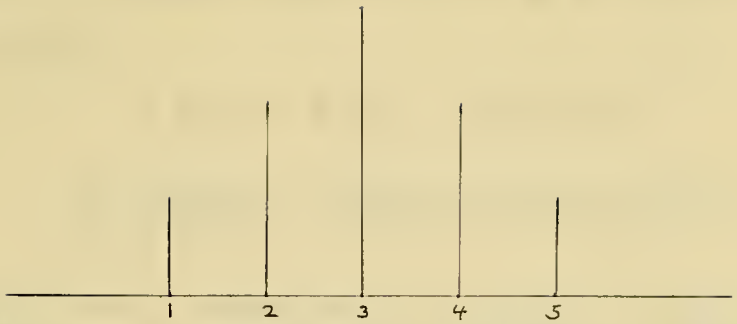


Figure 1.- Distribution of Cubes by Edge

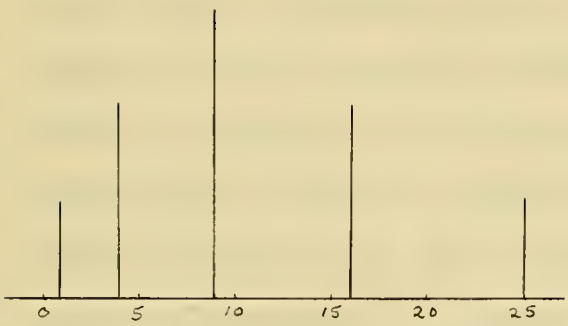


Figure 2.- Distribution of Cubes by Area

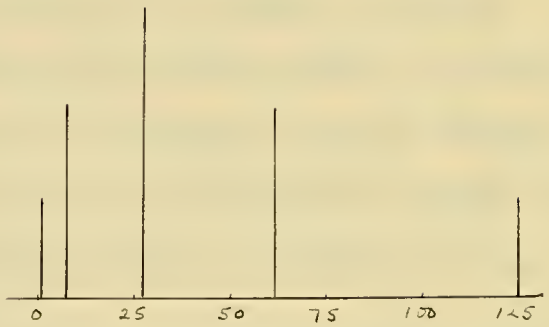


Figure 3.- Distribution of Cubes by Volume

It may be well to point out briefly that the inconsistencies in the above example are due to the relationships between the index variables;i.e.

$$v = e^3$$
$$a = e^2$$

It will be sufficient to note here that similar inconsistencies will arise whenever the relationship between the index variables is other than linear ($y = ax+b$). In the latter case the statistical constants will be merely affected by propor-

tionality factors. Thus when height is measured in inches and again in centimeters, the distributions will be similar and all constants easily converted by the simple linear relationship,

1 inch = 2.54+ centimeters

d. Indexing a Mental Characteristic

In mental measurement the method is to set a body of material before a child and elicit certain responses from him. These responses are then recorded and combined in various ways to produce what is known as a score. Moreover these responses exhibit considerable variety under different modes of administering the tests. Two boys took the Terman Group Test of Mental Ability which is administered under the plan of keeping the time constant. Their responses may be set forth briefly as follows:

	Score (Author)	Right	Attempted	Wrong	Accuracy
John	150	125	175	50	.71
Henry	150	140	150	10	.93

The problem of indexing here is not unlike that for the case of the cubes. Intelligence is the characteristic to be indexed, and this is possible by using the Rights, Wrongs, Attempts, or some combination of these index variables in the form of a score. It will be shown later that for the particular scale in question the author's score is a rather complicated function of the variables Right and Wrong, but for the present it is sufficient to note the possibility of such a characteristic

being indexed in several ways.

Questions immediately arise as to the best mode of indexing. Is it best to use only one of the index variables available, or to combine several of them, and if more than one variable is used how is the combination to be determined ? These questions are of most vital importance in mental measurement, arising in one form or another whenever a new test is devised and standardized. It may be pointed out at once that this study does not attempt any general solutions for these problems, but by analyzing a definite type of test material, aims to carry the investigation a little farther than is possible under incidental treatment in the construction of a particular scale.

Intelligence test material was chosen for two reasons. First, a large body of such data was available. Dr. F.S. Breed and Mr. E.R. Breslich made available their excellent data for three intelligence tests given at two grade levels in The University of Chicago High School. Mr. Guy Capps also turned over two thousand copies of the Terman Group Test of Mental Ability, Forms A and B, administered and repeated in a number of Kansas high schools. A second reason for using intelligence tests in this study was that the statistical constants resulting from such material are more stable than from any similar test data of which the writer is aware. The violent and often inexplic-

---

F.S. Breed and E.R. Breslich, Intelligence Tests and the Classification of Pupils, School Review, Vol. XXX (Jan. 1922) 51-66

able fluctuations in such constants as correlation coefficients and standard deviations for many of the well-known achievement tests for reading and arithmetic, would only complicate further the analysis attempted in a study of this kind. Such statistical stability, it is believed, is due for the most part to the length of the tests and to their careful construction.

Section 2 Groups and Data

a. The Groups Studied

For an analytical study of this type it is desirable that the groups studied should possess two important properties. They should be large enough to insure reasonable stability in the statistical constants; and second, they should be as homogeneous as possible. Representative samples from large populations are extremely difficult to obtain in test work but fortunately this property was not essential to the present study. Nevertheless for the largest group described below a rough sample of 100 out of 1000 cases was taken by combining three schools with approximately the same mean scores as the total group.

The group designated hereafter as Grade 7 consisted of 50 pupils from the University of Chicago Elementary School. Groups I High A and I High B included 50 and 60 pupils respectively from the University of Chicago High School. The 7th Grade pupils were doing part high school work, and were prepared to enter the regular first year classes the following year, there being no eighth grade in the Laboratory Schools.

Group I High B differed from I High only by the addition of 10 pupils in certain of the tabulations. While the groups described are undoubtedly select.they are unusually homogeneous as regards social status,training,experience with tests,and age. The largest group,I High C,consisted of 135 pupils from the Rolla ,Salem,and St. James public high schools of Kansas,selected as described above.

The age distributions for these groups are given in Table 3. The Kansas high school pupils were a year older than those in The University High School,while the latter were about a year and a half older than the 7th grade group. The distributions in each case present a fair degree of symmetry,the standard deviations increasing with the size of the group.

TABLE 3.-AGE DISTRIBUTIONS FOR THE GROUPS STUDIED

Age	Grade 7	I High A	I High B	I High C
19.0-19.5	..	..	..	1
18.5-18.9	..	..	..	1
18.0-18.4	..	..	1	2
17.5-17.9	..	..	..	6
17.0-17.4	..	..	1	8
16.5-16.9	..	2	2	13
16.0-16.4	..	3	3	13
15.5-15.9	..	2	2	17
15.0-15.4	..	6	7	18
14.5-14.9	2	9	10	21
14.0-14.4	2	11	13	16
13.5-13.9	6	10	12	9
13.0-13.4	5	4	5	6
12.5-12.9	16	3	3	3
12.0-12.4	10	2	2	..
11.5-11.9	8	..	..	1
11.0-11.4	1	..	..	..
Total	50	50	60	135
Mean	12.8	14.4	14.4	15.4
S.D.	.8	1.0	1.1	1.4

b. The Tests Used

Three intelligence tests were administered to the Laboratory School groups by Mr. Breslich during the year 1920-21. The scales used were the Otis Group Intelligence Scale, Form A, the Terman Group Test of Mental Ability, Form B, and the Chicago Intelligence Scale, Form B. These tests were all carefully scored by the writer for Attempts, Rights, ~~and~~ Wrongs, and Accuracy, as well as for the author's score.

The data from the Kansas schools consisted of the Terman Group Test of Mental Ability Form B, and the same tests Form B given the next day. From this group, therefore, it was possible to obtain the reliability coefficients. The large amount of labor involved in scoring each paper for author's score, total Attempts, Rights, Errors, and Accuracy, and carefully checking all of the work is largely responsible for limiting the sample to 135 cases when over 1000 were available.

In addition to the above data, school marks were obtained for 39 pupils in Grade 7. Yearly grades were obtained for Mathematics, English, and History. These marks were converted into a rough scale dividing pupils into seven catagories for purposes of correlation. It is the belief of the writer that this degree of accuracy in treatment is about all such data warrent, inasmuch as these marks are intended to give only a rough estimate of the

achievement in the various subjects.

Section 3 Test Administration and Scoring

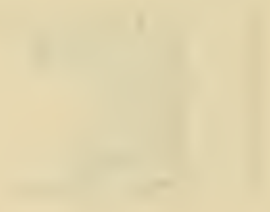
a. Primary and Secondary Index Variables

In administering and scoring a test the following variables must of necessity be taken into consideration directly or indirectly: Difficulty, Time, Attempts, Rights, Wrongs, and Omissions. This implies of course that the test material consists of a series of items where the responses may be scored right, wrong, or omitted. The six variables listed above will be referred to hereafter as Primary Index Variables and any function involving more than one of them as secondary index variables. Thus if Accuracy be defined as Rights divided by Attempts, such a score would be termed a Secondary Index Variable. Also in order to save space these variables will usually be denoted by the initial letter of each word, *i. e.*,

D = Difficulty
A = Attempts
T = Time
R = Rights
W = Wrongs
O = Omissions

b. The Difficulty Factor Eliminated

The problem of the difficulty of the various items in the tests is one which must be settled before proceeding farther. This problem is implicitly solved by the authors of the tests wherein each item is given an equal or point



value with all others. Two questions arise in this connection: are the items of equal difficulty, and if not should they be weighted to obtain an accurate score? It is a well known principle in the theory of index numbers that the longer the series the less the effect of differences in the weights of the individual items. Test scores are, of course, really index numbers. It may be readily conceded therefore, that the authors of intelligence scales consisting of so long a series of items, are quite justified in assigning equal weights to each item regardless of the better values which might be assigned on theoretical grounds. By way of justifying this assumption, an example will be given of a short series with considerable variation in weight from item to item. If differences in weight are not significant in so short a series, they will be even less so for a very long one. This method is that of the unfavorable case.

Test 2 of the Terman Group Intelligence Test consists of 11 items the response to each being a best answer appropriately checked. One hundred pupils were selected who had taken both Forms A and B a day apart. The test papers for Form A were then scored for errors and rough weights assigned in the usual way assuming a normal distribution of difficulty.

TABLE 3.-PER CENT FAILING AND WEIGHTS FOR TERMAN TEST 2

Item	1	2	3	4	5	6	7	8	9	10	11
Per cent Failing	8	19	29	30	26	24	16	47	65	32	55
Weight	2	3	4	4	4	4	3	5	7	4	5

Next by exceedingly tedious computation a weighted and an unweighted score for each pupil was obtained. Thus a boy responding correctly to items 1, 3, 7, 9, and 11 on the test received a weighted score of 21 and an unweighted score of 5, the number right. A correlation table was then made for weighted and unweighted score with a resulting coefficient of $r_{uw} = .972$. The standard error in estimating unweighted from weighted score, $S_u = .9$ .The correlation between the two unweighted forms of the test was then obtained, giving $r_{AB} = .597$ The standard error S_u in estimating unweighted form A from unweighted B was 4.8.

Thus the correlation between two forms of the same test is much lower than that between weighted and unweighted scores on the same form. Also the standard error of estimate of unweighted from weighted is about one-fifth that from form B. The weighting of the items then gives a degree of refinement considerably beyond the reliability of the test itself i.e. correlation of two forms. For 40 weighted and unweighted items of different material the writer obtained correlations of .998,.997, .995 with the corresponding reliability coefficients between .85 and .90. All of these results point to the conclusion that for a fairly long series weighting of the separate items is unnecessary. A complete solution of the problem would involve experimentation with series of various lengths, items of various difficulty, and populations of different size. Such exhaustive treatment is clearly beyond the scope of this study. It may be finally pointed out that a number of recent achievement tests have appeared first with weighted items and later with weights dropped when it

was realized what a slight difference these made in the resulting scores.

C. Methods of Administering the Tests

With the factor of difficulty eliminated the scoring or indexing problem is greatly simplified. The plan may now be described as the method of Unit Responses, the response to each item being scored as a unit point. Furthermore if omissions be neglected or counted as errors, the remaining primary index variables are reduced to Time, Attempts, Rights, and Wrongs with the relationship,

$$A = R+W$$

if omissions be counted as errors. This assumption will be made in a subsequent discussion. The number of omissions occuring in the tests used was negligible. It thus a pears that four variables, T, A, R, and W will have to be studied, the last three not being independent. Also all scoring formulae or secondary index variables will be functions of these four primary indexes.

Two plans for administering such tests are possible. One plan is to fix the number of Attempts allowing Time, Rights, and Errors to vary, while the second method is to fix the Time, allowing Attempts, Rights, and Wrongs to vary. In the last analysis then, only three index variables need to be considered, the fourth being arbitrarilly fixed by the plan of administering the tests. According to the first scheme outlined, one allows all of the children to

finish the test thus keeping Attempts constant. The time is then recorded by stop watches or clock device and Rights and Wrongs obtained from the papers. By the second plan a fixed time limit is set for all the pupils, Attempts Rights, and Wrongs being then scored on the test papers. The second method is obviously simpler than the first, and is now followed in the great majority of tests of all kinds. Certain tests, it is true, neglect actual time but these are not considered here. The material used in this study is all administered under the plan of fixing time giving the three primary index variables A,R, and W.

d. Authors' Plans of Scoring

For the three intelligence tests described above, the authors have set forth scoring formulae for each test of the battery. These formulae are obviously expressed as functions of the primary variables A,R, and W, while relationship A = R+W makes it possible to set down the equations in terms of any two of the variables. In the following tabular scheme, therefore, R and W have been employed throughout.

TABLE 4.- AUTHORS' SCORING FORMULAE ON THE THREE SCALES

Scale	Test									
	1	2	3	4	5	6	7	8	9	10
Otis	R	R	R-W	R	R	R	R	R	R	R
Terman	R	2R	R-W	R	2R	R-W	R	R-W	R	2R
Chicago	$R-\frac{1}{2}W$	2R	$2(R-\frac{1}{2}W)$	R	2R	—	—	—	—	—

It will be seen from table 4 that two of the scales consist of ten tests each while the third includes only five. Considerable variation in the scoring formulae is also to be noted. The otis Scale has 9 tests scored by R and only one by R-W; Terman, on the other hand, scores four tests by R, 3 by 2R, and 3 by R-W. For the Chicago Scale two new formulae appear, R- $\frac{1}{2}$ W and 2(R- $\frac{1}{2}$ W). All of the above formulae are clearly special instances of the general linear form

$$S = a(R+bW)$$

where a and b are constants.

The simplicity of the Otis' scoring formulae immediately raises the question as to the advisability of scoring the other scales by the same method. This problem is discussed in the following sections. Another question concerns the formulae R-W and R- $\frac{1}{2}$W. These formulae are employed for material of the True-False type and three-choice variety, on the supposition that they correct for the element of guessing for such tests. This problem will also be taken up in some detail in later discussion. Finally the doubling of scores for individual tests is a point that needs consideration. So far as the writer is able to determine, this doubling was effected because the author of the test felt such tests to be worth about twice as much as others, or wanted to increase the total points possible to some convenient number. The point at issue is the same as that

between weighted and unweighted items. A graduate student made a study of the Chicago Scale regarding this problem and found that the weighting of the three tests as indicated in Table 4 affected the correlation but slightly, the coefficient between weighted and unweighted scores being .98.

It thus appears that Rights are the basis of most of the scoring formulae employed on these tests and that other forms have been used to correct Right responses for guessing or to weigh the whole test because of the relative importance in the battery making up the scale.

Section 4 Methods Employed

The general method of this study will first be to analyze the interrelationships of the index variables involved, and then to set up certain criteria of good indexes and attempt to evaluate the variables in terms of these. Obviously this is an indirect approach and it must necessarily be so from the nature of the problem. The actual technique employed will involve a considerable amount of correlation, for this is the best method of studying the relationships between index variables from tests. For the case of the cubes described above the index variables

were functionally related, i.e.

$$\text{volume} = (\text{edge})^3$$

Such functionality can only be approached by empirical data, the correlation coefficient giving the most convenient approximation for linear functions. Thus if score and Right. are correlated to the extent of .98 with regression linear, there is a very close approximation to the functional relationship S = KR. The chief advantage of correlation is that it gives a numerical estimate of the closeness of such relationships or approach to linear functionality.

Index variables will be analyzed by batteries and by single tests. Age, school marks, and other intelligence tests will be used as criteria against which to check the various formulae.

All of the calculations below have been preformed by the writer and have been checked with care. Correlations were obtained by the usual product moment method with a specially designed correlation form. This was found to cut down the labor of calculation very materially especially when "batteries" of coefficients were required. Blakeman's test for linearity was applied in a few cases with the result that the writer believes the great majority of the tables exhibited sufficient linearity so as not to reduce the correlation beyond the limits of probable error.

Part I will be concerned with an analysis of the five index variables by whole scales. By the indirect method of correlation, the relative merit of the indexes will be determined. Part II will involve a similar analysis of the individual tests of the scales. In Part III certain form-

ulae will be developed and their validity and reliability determined in mathematical terms.

Part I Analysis of the Index Variables by Whole Scales

A.- Comparative Validity and Reliability of the Indexes

The most direct approach to the problem of indexing will be to compare the author's plan of scoring as exhibited in Table 4 with the simpler plans of counting merely Attempts, Rights, Errors, and determining Accuracy for whole scales. This procedure will reveal the relative merit in reliability of these simpler primary variables when several tests are pooled to give a scale score. The term "Scale Score" is here employed to distinguish it from test score which will be used to denote the score on one of the components making up the total scale. Thus the Otis Scale is made up of 10 component tests.

Five types of comparison will be made in evaluating the index variables for general reliability. These include:

a. Relationships between index variables on the same scale
b. Relationships between scales by the same index variables
c. Reliability of a scale by different index variables
d. Correlations with age
e. Correlations with school marks

Section 5 Relationships between Index Variables on the Same Scales

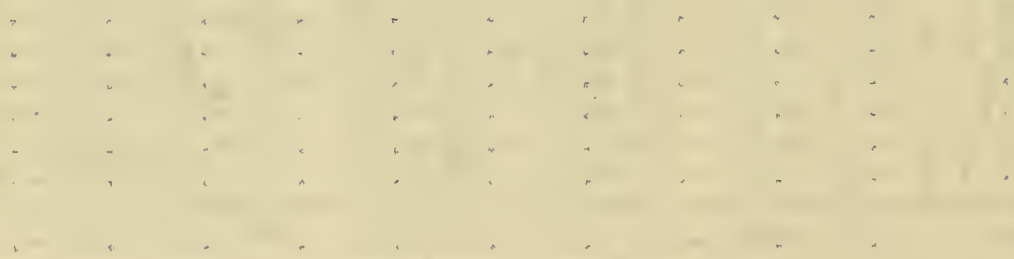
As pointed out above the relationships between the va-

ibles cannot in general be functional for empirical data such as those obtained from mental tests. Nevertheless it will be valuable to discover the closeness of linear relationship as indicated by the correlation coefficient. These results are set forth in Table 5. All of the correlations were computed on total Attempts, Rights, etc. for the entire scales. The agreement from group to group and scale to scale between correlations for the same two variables is striking and may be viewed with pardonable satisfaction by one who has viewed many inexplicable differences for other types of tests and smaller groups.

TABLE.5.- CORRELATIONS BETWEEN VARIABLES FOR THE SAME SCALES

Scale and Group	Pairs of Variables Correlated									
	S×A	S×R	S×W	S×$\frac{R}{A}$	A×R	A×W	A×$\frac{R}{A}$	R×W	R×$\frac{R}{A}$	W×$\frac{R}{A}$
Otis 7	+.72	+.99	-.52	+.73	+.73	+.21	+.13	-.51	+.75	-.91
Otis I B	+.67	+.99	-.59	+.81	+.67	+.30	+.02	-.57	+.77	-.93
Ter.7	+.59	+.96	-.45	+.66	+.70	+.41	-.16	-.30	+.53	-.94
Ter.I B	+.59	+.96	-.67	+.33	+.74	+.14	+.10	-.51	+.73	-.93
Chi. 7	+.45	+.90	-.54	+.69	+.69	+.31	-.10	-.39	+.59	-.93
Chi. I B	+.54	+.95	-.71	+.33	+.62	+.19	+.03	-.65	+.70	-.94
Mean	+.59	+.96	-.58	+.70	+.69	+.34	.00	-.49	+.69	-.93

for example the correlation between authors' score and total Attempts for Otis Grade 7 is .72 ±.05, while for Otis I High B it is .67 ±.05. By the formula $P.E_{a-b} = \sqrt{(P.E_a)^2 + (P.E_b)^2}$ the difference between the two correlations may be written in the form diff.=.05±.07 , indicating that a difference of .05 is unsignificant. To be significant such a difference would have to be 3 times its P.E. Similar comparisons between groups show that nearly all diff-



ences may be accounted for by the fluctuations in sampling. Moreover if tests made on the same group with different scales be considered as independent samples, most of the inter-scale differences are also insignificant.

The table yields some very interesting results. In the second column it will be observed that the correlations between S and R are very high, the mean of the six coefficients being +.96. The extremely high correlation of .99 for the Otis scale is due to the fact that only one of the tests in this scale is not scored directly by Rights. The above result, therefore, raises a question as to the advisability of scoring this lone test differently from the rest. Moreover the two correlations of .96 for the Terman Scale indicate that even with a fairly complicated system of scoring, the agreement between authors' score and total Rights is extremely close. With still more complicated formulae in the Chicago Scale the correlations between S and R are again very high. For whole scales, then, a mere enumeration of the total number of right responses gives a result very nearly proportional to that obtained by the use of various formulae for the individual tests making up the scale. It is to be noted, however, that the above result appears to be valid when single tests are pooled to give a total score. For single tests, changes in the scoring formulae have marked effect upon the correlations with criteria as will be shown later.

Accuracy furnishes the next highest correlation with Score, the mean coefficient being +.76, while A and W come next with correlations of +.59 and -.58 respectively. If Score be adopted as a criterion, therefore, the best index variables in order would be Rights, Accuracy, and Attempts or Errors. It is rather surprising that merely counting the Attempts or Errors on these tests gives so good an index of intelligence.

In the last column of the table a very high and consistent negative correlation is found between W and $\frac{R}{A}$. This means that a pupil who makes a great many errors is very inaccurate as measured by the index $\frac{R}{A}$, or that such an index variable is an exceedingly good measure of what is ordinarily understood by Accuracy. It is of some interest to note that the above correlation becomes -1.00 when the number of attempts is fixed and the relationship

$$A = R + W$$

still obtains. In this case the functional relationship

$$R = -W + \text{Const}$$

holds strictly, and this implies perfect negative correlation as is shown more fully in a section below.

If the index variable A be used as a measure of speed a number of interesting relationships are brought out by the remaining coefficients. Speed and accuracy are evidently uncorrelated, the highest correlation between these two variables being -.16 ± .09 which is insignificant because it is not even twice its probable error in amount. Moreover

the differences in sign are such as to give a mean correlation for the column of less than .01. It will be further noted that R is correlated with A to the extent of +.69 on the average, while W with A gives a mean correlation of +.24. An average correlation of +.59 between S and A was previously noted. These results indicate that the way to get a high intelligence score is to work fast. By working fast one is likely to make more mistakes, but he is much more likely to get more items right and make a higher score than if he worked more slowly. Intelligence tests have frequently been called "alertness" tests. The above findings indicate that with considerable appropriateness they might also be termed "speed" tests. The assumption thus far in determining general reliability has been that the authors'score is the best index of intelligence. If, on the other hand, Accuracy had been assumed to be the best index, the speed factor would have been eliminated, there being no general tendency as indicated by the zero correlation for a pupil to get a high or low accuracy score by changing his speed. It will also be recalled that Accuracy was highly correlated with Score (+.76) so that $\frac{R}{A}$ as an index variable is ~~inferiorly~~ fairly consistent with S without being unduely influenced as is the latter variable by the undesirable speed factor.

The average correlation of -.49 between R and W means that the more items the pupil gets right, the fewer he is

likely to get wrong. If Attempts are constant the above correlation becomes -1.00 as in the case of Wrongs and Accuracy. Finally the variables R and $\frac{R}{A}$ are found to correlate on the average +.63, a coefficient which becomes +1.00 when Attempts are constant. (Theorem 3 Appendix)

Section 6. Relationships between Scales by the Same Index Variable

The relationships between index variables on the same scale have been discussed in the preceding section, with the result that the general order of merit with score as a criterion is Rights, Accuracy, and Attempts or Errors. The variable R possesses a simplicity which considered in connection with its close agreement with S suggests that it might well be substituted for that variable in indexing by batteries of tests. The index $\frac{R}{A}$ was found to be in close agreement with S and R and to possess the advantage of being unaffected by speed.

The next procedure will be to evaluate the index variables indirectly by determining the correlations between pairs of scales indexed by the same variables. The closeness of this correspondence will give a measure of the effectiveness of the particular mode of indexing. The former comparison was inter-variable, the present is inter-test. Table 6 gives these results for groups I High B and Grade 7. Comparison of this table with Table 5 reveals the fact that the correlations in the former are much more nearly the same size. All of the coefficients are significant, while there are few of the differences between any

two which may not be attributed to sampling by the usual method. The means for the columns and rows of the table bring out the stability more clearly inasmuch as the deviations from these are in most cases very slight.

The really surprising results exhibited by this table are indicated in the column of means for the various indexes. Here it appears that with inter-scale correlation as a criterion, S, R, $\frac{R}{A}$, and W are all about equally good for purposes of indexing, while A is somewhat poorer than the rest. To discover almost as high a correlation between two intelligence scales by merely tabulating total errors as by using the author's score or total Rights is at first somewhat surprising. It is a closer agreement than might be expected from the average correlation of -.58 between S and W from Table 5. Inter-test correlation gives a measure of the extent to which two tests agree in measuring the same characteristic. From Table 6 it appears that this agreement is about equally close when any of the four index variables is employed, the general order of merit being indicated in the table as S, R, $\frac{R}{A}$, W, and A. This order, it will be recalled, is in harmony with that found in the preceding section. It may finally be pointed out that the lack of variability among the coefficients strongly suggests that combining these index variables as in the case of S will give but slightly better index when batteries of tests such as these are employed.

TABLE 6.-INTER-SCALE CORRELATIONS BY FIVE INDEX VARIABLES

Index Variable	Pairs of Scales and Groups						Mean	Order
	Term. × Otis		Term. × Chic.		Chic. × Otis			
	El.	H.S.	El.	H.S.	El.	H.S.		
Score	+.72	+.83	+.68	+.76	+.76	+.79	+.76	1
Attempts	+.66	+.72	+.50	+.46	+.50	+.53	+.56	5
Rights	+.73	+.82	+.59	+.71	+.78	+.83	+.74	2
Wrong	+.72	+.83	+.60	+.69	+.65	+.65	+.69	4
Accuracy	+.70	+.88	+.61	+.77	+.66	+.74	+.73	3
Mean	+.71	+.82	+.60	+.68	+.67	+.71	+.70	

Section 7 The Reliability of a Scale by Different Index Variables

Another method of studying the relative merit of the different index variables is to obtain the reliability coefficients for a scale under each of the indexes. Group I High C was used for this purpose. It will be recalled that this group consisted of 135 pupils who took Form B of the Terman Scale and Form A of the same test on the following day. The reliability coefficients in this case are given by the correlations between the variables on the two forms of the scale. These correlations are given in Table 7 with the contingency tables presented in the appendix.

TABLE 7.- RELIABILITY COEFFICIENTS FOR THURIAN FORMS A AND B ON GROUP I HIGH C

Variables	Reliability Coefficient
Score	+.908 ±.010
Attempts	+.684 ±.032
Rights	+.896 ±.011
Wrongs	+.737 ±.027
Accuracy	+.841 ±.017

It is at once apparent that the author's score has the highest reliability with a coefficient of .908. A glance at the corresponding correlation table in the appendix will give this result more meaning. Here the linearity of regression is at once apparent, and the remarkable agreement exhibited in graphical form. Close correspondence of this type appears to the writer as one of the most significant achievements of standardized tests. Just what the tests do measure or index is often ambiguous, but to index any mental characteristic with such a high degree of reliability is in itself a most noteworthy achievement. It should also be borne in mind that all such reliability coefficients (and indeed all such correlations) depend upon the group. Selection will in general tend to reduce such correlations while heterogeneity due to such factors as age will tend to increase it.

Returning to the remaining coefficients in the above table, we find that the order of reliability for the five variables is, S, R, $\underset{A}{R}$, W, and A. This is precisely the order obtained by the method of inter-test correlation as shown in Table 6.

If the symbol Π_{xx} be employed to denote the reliability coefficient for a test indexed by the variable X, the differences between the correlations in Table 7 may be exhibited as follows:

$\Pi_{ss} - \Pi_{AA}$	=	$+.224 \pm .034$	difference is significant
$\Pi_{ss} - \Pi_{RR}$	=	$+.012 \pm .015$	difference is insignificant
$\Pi_{ss} - \Pi_{ww}$	=	$+.171 \pm .029$	difference is significant
$\Pi_{ss} - \Pi_{\frac{RR}{AA}}$	=	$+.067 \pm .020$	difference is significant

The formula used for calculating the probable errors of the differences is,

$$P.E_{a-b} = \sqrt{(P.E_a)^2 + (P.E_b)^2}$$

The small differences between the reliability coefficients for S and R is insignificant, while the remaining differences are sufficiently large in comparison with their probable errors to be significant. Thus from the standpoint of reliability the Terman Scale is indexed equally well by author's score or total rights, and next best by accuracy. A reliability coefficient of .74 is generally considered high, and it is remarkable that total errors on the two forms should correlate to such an extent. The difference between Π_{ss} and Π_{ww}, however, is over five times its probable error so that the reliability of errors as an index is significantly less than for score. Similarly Attempts furnish a much less reliable index than Score, Rights, or Accuracy.

In this part of the study no attempt is made to analyze the individual tests making up the scales. Nevertheless it is interesting to note from Table 4 that three

of the Terman tests are scored R-W, three All and four by R alone. The difference in the reliability coefficient r_{ss} under this plan, and r_{RR} by counting merely rights is $+.021 \pm .015$ as shown above. Formulae of the type 3R and R-W, thus appear to have no effect on the reliability of the total score, and their use for such scales is open to question.

Section 8 Correlations with Age

The age factor is always of interest when studying test results. Table 3 which gives the age distributions in half-year intervals, shows ranges for the various groups from four to eight years. In Table 8, the correlations between age and the four index variables are given for Grade 7 and I Hig B. A consistency in these coefficients is at once apparent. The mean correlation of -.41 between age and score indicates that the younger pupils are brighter than the older ones within the same grade group. Similarly the mean correlations -.38, -.29, -.36, and +.20 show that the younger pupils get more items right, are more accurate, are speedier, and make fewer errors than the older children.

TABLE 8.- CORRELATIONS BETWEEN AGE AND INDEX VARIABLES

Scale and Group	Age with				
	Score	Attempts	Rights	Wrongs	Accuracy
Otis 7	-.41	-.37	-.37	+.06	-.24
Otis IB	-.40	-.27	-.39	+.27	-.32
Terman 7	-.47	-.22	-.40	+.17	-.27
Terman IB	-.39	-.21	-.37	+.28	-.34
Chicago 7	-.39	-.21	-.35	+.20	-.33
Chicago IB	-.39	-.30	-.39	+.24	-.26
Mean	-.41	-.26	-.38	+.20	-.29

The superiority of the younger child is evident, therefore, no matter which index is employed for the tests.

When arranged according to the size of the correlation with age the order of the index variables appears from the means as, S, R, $\frac{R}{A}$, A, and W. The three variables Score, Rights, and Accuracy retain the order found in the previous sections. Moreover the mean correlations for score and rights with age are very nearly the same, so that the superiority of S over R as an index is again slight if any. The usual sampling formula reveals no difference in the correlations that is of statistical significance.

Similar correlations are given in Table 9 for the largest group. The coefficients, though somewhat smaller are in harmony with those of the preceding table. The decrease in size is probably due to the greater range in age.

TABLE 9.- CORRELATIONS OF INDEX VARIABLES WITH AGE FOR TERMAN SCALE FORM B

Group	Age with				
	Score	Attempts	Rights	Wrongs	Accuracy
Grade 7	-.47±.07	-.32±.09	-.40±.08	+.17±.09	-.27±.09
I High B	-.39±.08	-.21±.09	-.37±.08	+.28±.09	-.34±.08
I High C	-.16±.06	-.14±.06	-.16±.06	+.03±.06	-.12±.06

By increasing the range through several grades the correlation becomes positive. In fact heterogeneity or lack of selection appears to have a curious effect on correlations with age. For children of exactly the same age the coefficient is of course zero; when the range is increased to several years as in the typical grade group, the correlation is negative. As the range in age increases, the coefficient approaches zero again, and finally passes through this value to positive values of considerable size if several grades are pooled to give a long range. The theoretical curve for the correlation coefficient will thus have an appearance resembling that in Figure 4. The negative correlation for the age interval OA has been accounted for by the appearance of older retarded children in grade groups. This explanation, while plausible, does not seem satisfactory for groups such as Grade 7, which is usually free from children of this type. The positive correlation increases from A as the age span is lengthened.

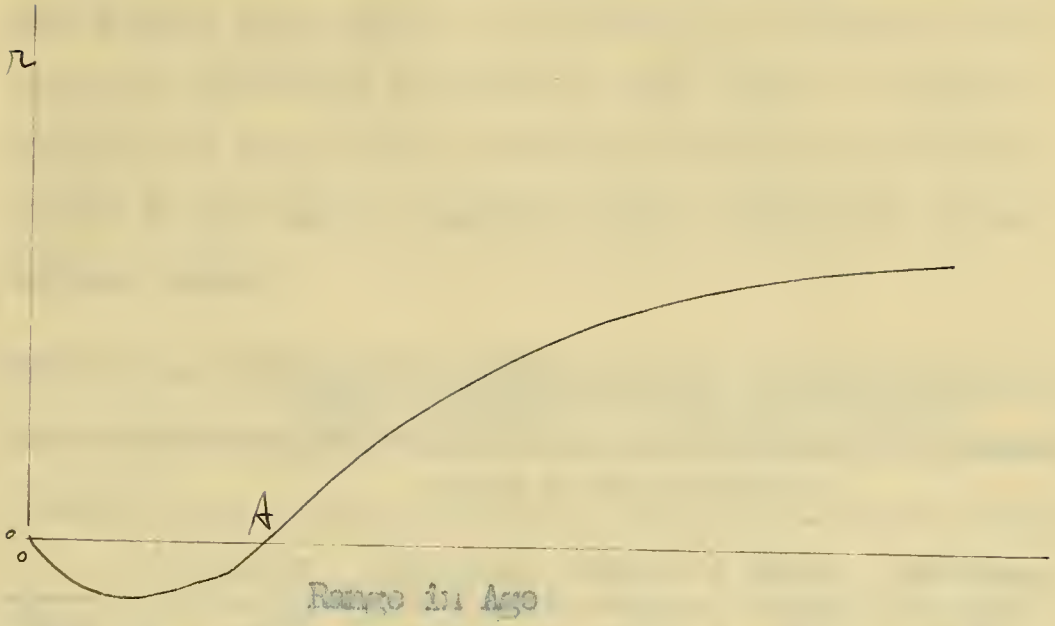


Fig. 4 Theoretical curve for correlation with age

Section 3 Correlations with School Marks

While school marks are obviously inaccurate estimates of the ability or achievement of pupils, nevertheless they may serve as a useful device against which to check test results. Their relative inaccuracy has been grossly exaggerated. Burt,[a] Proctor,[b] Kelly,[c] and others have shown that the predictive value of such marks is often as high as

that from intelligence or achievement tests. Moreover in the present comparisons the question of accuracy is not of great importance inasmuch as each index is checked against the same school grades. Whatever unreliability exists in the marks, therefore, will affect all correlations alike.

TABLE 10.- CORRELATIONS BETWEEN MARKS IN ENGLISH AND THE INDEX VARIABLES ON THE THREE SCALES

Scale	Marks in English with				
	Score	Attempts	Rights	Wrong	Accuracy
Otis	+.56	+.36	+.57	-.29	+.46
Terman	+.63	+.30	+.56	-.32	+.46
Chicago	+.52	+.27	+.53	-.38	+.48
Mean	+.57	+.31	+.55	-.33	+.47

TABLE 11.- CORRELATIONS BETWEEN MARKS IN HISTORY AND THE INDEX VARIABLES ON THE THREE SCALES

Scale	Marks in History with				
	Score	Attempts	Rights	Wrong	Accuracy
Otis	+.53	+.17	+.54	-.45	+.55
Terman	+.63	+.24	+.60	-.38	+.53
Chicago	+.48	+.17	+.41	-.37	+.47
Mean	+.55	+.19	+.52	-.40	+.52

TABLE 12.- CORRELATIONS BETWEEN MARKS IN MATHEMATICS AND INDEX VARIABLES ON THE THREE SCALES

Scale	Marks in Mathematics with				
	Score	Attempts	Rights	Wrong	Accuracy
Otis	+.60	+.45	+.61	-.24	+.40
Terman	+.54	+.32	+.72	-.26	+.36
Chicago	+.47	+.29	+.36	-.17	+.27
Mean	+.54	+.35	+.53	-.22	+.34

Correlations of the five index variables with English, History, and Mathematics are given in Tables 10, 11, and 12 respectively. Inspection of the average correlations for these tables shows a close agreement for the three subjects studied. The coefficeints for English are highest, History next, and Mathematics lowest, but the differences are slight. In all three tables Score has the highest correlation with school marks. The next highest correlations in order are Rights, Accuracy, Wrongs, and Attempts. This is precisely the order found in the section on reliability. Thus if school work be measured by marks the relative merit of the various indexes for prediction is S, R, $\frac{R}{A}$ W, and A. As in the preceding sections, correlation involving S and R are more nearly equal than the others.

Section 10. Summary for Reliability of Indexes for Whole Scales

For whole scales consisting of batteries of tests, the authors' formulae appear to be slightly superior to total Rights as as index. Table 13 gives the average correlations and differences in favor of S (absolute values considered)

TABLE 13.- SUMMARY OF CORRELATIONS FOR SCORE AND RIGHTS

Variables Correlated	Average Coefficient	Difference in Favor of S
Score and Rights	+.96	
Scales Indexed by Score	+.76	
Scales Indexed by Rights	+.74	.02
Terman Forms B and A by S	+.91	
Terman Forms B and A by R	+.90	.01
Age with Score	-.41	
Age with Rights	-.38	.03
Marks with Score	+.55	
Marks with Rights	+.53	.02
Total Difference		.08

The extreme simplicity of scoring by Rights, however, would seem to more than outweigh the slight advantage in favor of more complicated formulae.

Accuracy has been shown to have the peculiar advantage of being unaffected by speed, and at the same time to possess high reliability. The summary correlations in favor of Score are shown in Table 14. The total differ-

TABLE 14.- SUMMARY OF CORRELATIONS FOR SCORE AND ACCURACY

Variables Correlated	Average Coefficient	Difference in Favor of S
Score and Accuracy	+.76	
Scales Indexed by Score	+.76	
Scales Indexes by Accuracy	+.73	.03
Terman Forms B and A by S	+.91	
Terman Forms B and A by R	+.84	.07
Age with Score A	-.41	
Age with Accuracy	-.29	.12
Marks with Score	+.55	
Marks with Accuracy	+.44	.11
Total Difference		.33

ences in favor of S are shown in Table 14. The total differences in favor of S indicates that Accuracy is somewhat less satisfactory than R according to the criteria employed. Moreover it is a more involved complicated index than R, but not so involved as S.

The results for Errors are presented in Table 15. The general merit of Wrong as an index is less than that of the preceding variables. Errors, however, have a surprisingly high reliability and are utilized to advantage in formulae discussed in the following sections.

TABLE 15.- SUMMARY OF CORRELATIONS FOR SCORE AND ERRORS

Variables Correlated	Average Coefficient	Differences in Favor of Score
Score and Wrong	-.58	
Scales Indexed by Score	+.76	
Scales Indexed by Wrongs	+.69	.07
Terman Forms B and A by S	+.91	
Terman Forms B and A by W	+.74	.17
Age With Score	-.41	
Age With Wrongs	+.20	.21
Marks With Score	+.55	
Marks With Wrongs	-.32	.23
Total Difference		.68

Attempts, which are frequently used as an index for tests , appear to have the least merit of any of the variables discussed. Table 16 gives the averages and differences as in the above tables. The total absolute difference in favor of S is greater than for any of the preceding variables.

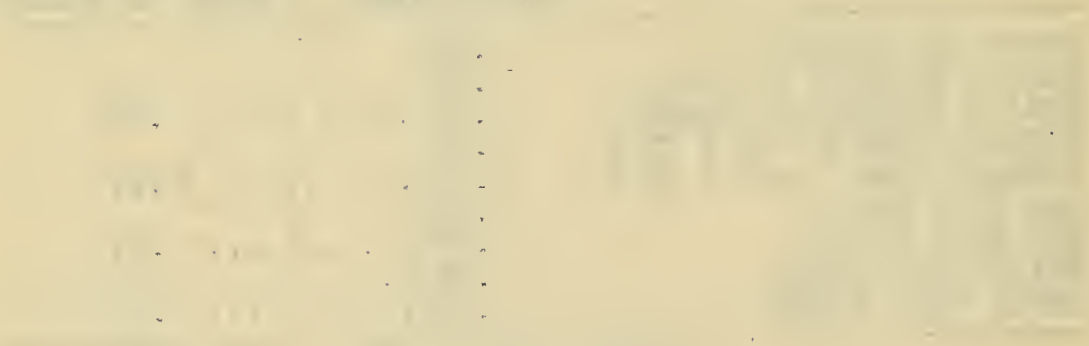
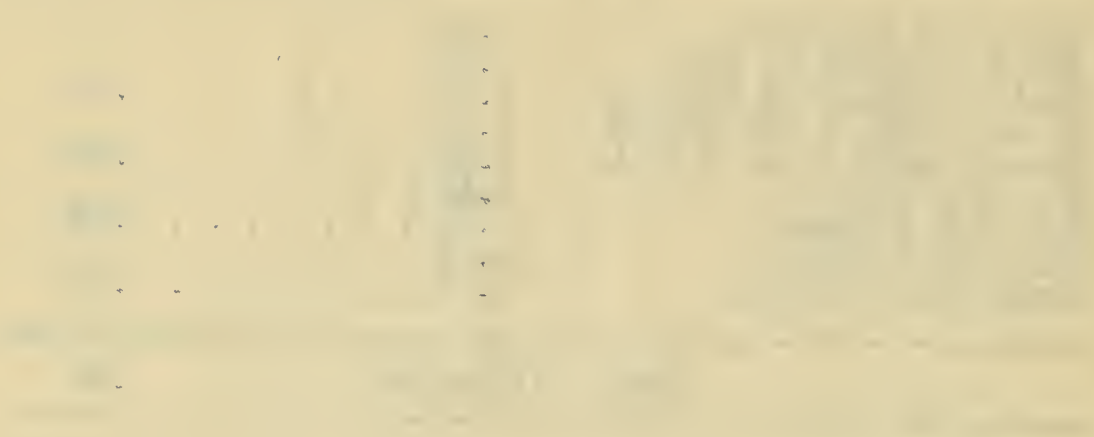


TABLE 16.- SUMMARY OF CORRELATIONS FOR SCORE AND ATTEMPTS

Variables Correlated	Average Coefficient	Difference in Favor of Score
Score and Attempts	+.59	
Scales Indexed by Score	+.76	
Scales Indexed by Attempts	+.56	.20
Terman Forms B and A by S	+.91	
Terman Forms B and A by A	+.68	.23
Age with Score	-.41	
Age with Attempts	-.26	.15
Marks with Score	+.55	
Marks with Attempts	+.28	.27
Total Difference		.85

B.- The Discriminative Capacity of the Indexes

In addition to the general reliability of an index, another valuable property of such a variable is the extent to which it makes possible discrimination between individuals and between groups when real differences exist. A test which reveals too narrow a range for a given group fails to discriminate between the individuals of that group. Such undistributed score is a defect in the test or in the mode of indexing. Similarly a test or mode of indexing which fails to discriminate between groups is defective if the characteristic is in reality different in type from group to group. Thus a test which shows all individuals in Grade 5 to possess the same ability, and at the same time reveals



no difference between mean scores for Grade 5 and 6 is lacking in individual and in group discrimination. The fundamental assumption is, of course, that such individuals and groups do vary and that failure to detect the variations lies in the particular mode of indexing the trait in question.

Section 11 Capacity of the Indexes to Discriminate between Individuals

Discrimination between individuals of a group is best studied by means of frequency distributions. In the present study, however, such an elaborate method as this is unnecessary inasmuch as intelligence tests are of sufficient length to give a fairly good spread for all indexes. The distributions for S, R, A, W, and $\frac{R}{A}$, in the Appendix are typical of those for all three intelligence scales. The standard deviation for these variables are given in Table 17.

TABLE 17.- STANDARD DEVIATIONS FOR TERMAN FORMS A AND B WITH GROUP I HIGH C

Variable	σ_B (first)	σ_A (second)	$\sigma_B - \sigma_A$	Diff÷P.E.diff.
Score	33.79±1.39	30.91±1.27	2.88±1.88	1.5
Attempts	24.12±0.99	20.63±0.85	3.49±1.30	2.7
Rights	26.95±1.11	23.54±0.97	3.41±1.47	2.3
Wrongs	19.73±0.81	19.66±0.81	0.07±1.15	0.1
Accuracy	0.13±0.01	0.12±0.01	0.01±0.01	1.0

There is some evidence that there is less variability in performance on the second trial (Form A) than on the first (Form B). This is incidentally a bit of evidence to the

effect that equal practice for a group of pupils tends to bring them more closely together about a central type, a result contrary to that held by some psychologists. The differences, however, are slight, although in one direction, and the two forms of the test may not be equivalent for this purpose. The result is then merely suggestive.

The standard deviations for A,R, and W in Table 17 admit of direct comparison inasmuch as they are all expressed in point or response units. The order of discriminative capacity for these variables is then R,A, and W. The indexes S and $\frac{R}{A}$ are expressed in different units and hence may not be compared with the rest. Considered on the point basis however, author's score has the greatest capacity for discrimination between individuals on account of the weighting and formulae involved. The standard deviations for Grade 7 and I High B are also given in Table 18. The results agree with those of the preceding table.

TABLE 18.- STANDARD DEVIATIONS FOR THE THREE SCALES

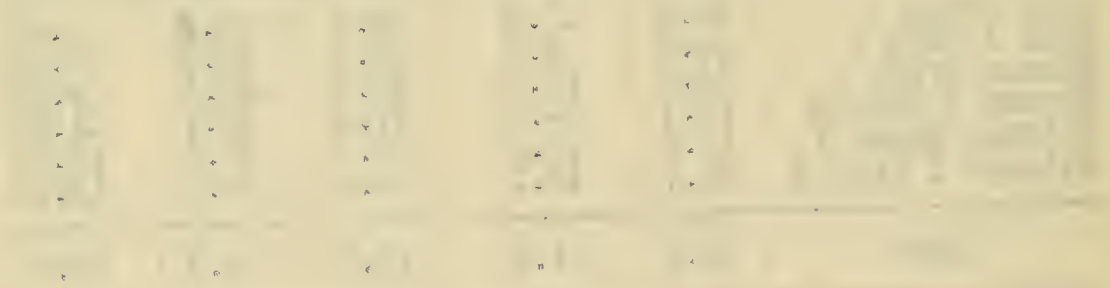
Scale and Group	Standard Deviations for				
	Score	Attempts	Rights	Wrong	Accuracy
Otis Grade 7	19.91	18.00	19.92	13.48	0.08
Otis I High B	22.40	18.02	21.55	17.02	0.08
Terman Grade 7	23.52	19.17	18.52	12.46	0.08
Terman I High B	26.87	18.39	22.34	14.87	0.09
Chicago Grade 7	10.52	6.20	6.15	4.60	0.09
Chicago I High B	12.03	5.80	8.24	6.50	0.10

In order to study the variability of a group by a statistical measure independent of the units employed, Pearson's Coefficient of Variation, $V = \frac{100\ S.D.}{M}$, was employed. The results for two groups appear in Table 19. It is at once apparent that while V is independent of the units employed it may nevertheless lead to results which are confusing. The largest coefficients of variation are for W, an index which might readily be supposed to furnish the least variability. The result is brought about by the relatively large standard deviation of W (Table 18) and the relatively low mean (Table 20, below).

TABLE 19.- COEFFICIENTS OF VARIATION FOR THE THREE SCALES

Scale and Group	Coefficients of Variation for				
	Score	Attempts	Rights	Wrong	Accuracy
Otis Grade 7	14.2	10.2	14.3	36.2	9.86
Otis I High B	14.8	9.5	14.2	45.6	10.35
Terman Grade 7	18.8	13.8	16.6	45.5	9.85
Terman I High B	18.7	11.5	17.4	47.7	11.37
Chicago Grade 7	19.5	11.7	14.6	43.0	10.71
Chicago I High B	21.7	10.1	18.8	48.1	13.18
Mean	17.8	11.1	16.0	44.3	10.89

The coefficient of variation, depending as it does upon the position of the distribution on the scale, is likely to give a very misleading result for distributions such as these above, and should in general be avoided for comparisons of this type.



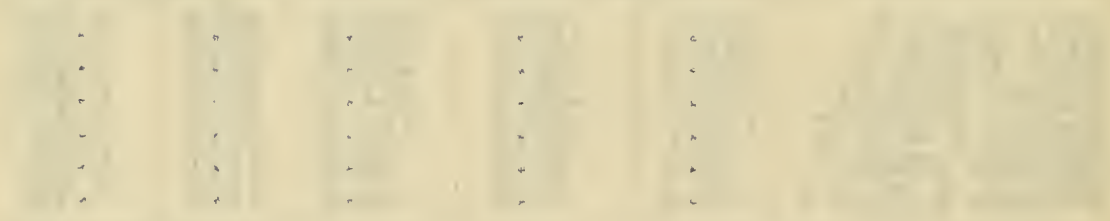
Section 12 Capacity of the Indexes to Discriminate between Groups

Table 20 gives the means on the three scales for Grade 7 and for I High B. It is at once evident that the second group has the higher mean for nearly all of the indexes. Accuracy, however, appears to be nearly consistent for all three scales and for both groups. From the standpoint of discrimination, therefore, this index is of little value. The correlation tables in the Appendix show a considerable spread for Accuracy while the constants from Table 17,18, and 19 indicate the extent of this variability.

TABLE 20.- MEANS FOR THE THREE INTELLIGENCE SCALES

Scale and Group	Means for				
	Score	Attempts	Rights	Wrong	Accuracy
Otis Grade 7	139.8	177.0	139.6	37.2	0.80
Otis I High B	151.3	189.3	152.0	37.3	0.80
Terman Grade 7	125.0	138.8	111.4	27.4	0.80
Terman I High B	143.7	159.5	128.3	31.2	0.81
Chicago Grade 7	54.0	52.8	42.1	10.7	0.80
Chicago I High B	55.4	57.2	43.7	13.5	0.76

Individuals within a group, then, differ considerably in accuracy. When the above inter-group comparison is made however, Accuracy is found to be relatively constant. These results indicate that the growth curve for accuracy ordered relative to age will be relatively flat in comparison with ordinary score. This whole matter will be



fully treated by the writer in a forthcoming article on Growth Curves under Different Modes of Indexing.

In order to bring out such inter-group differences more clearly they are presented in full in Table 21. The quantity $\frac{D}{P.E._{diff}}$ denotes the inter-mean difference divided by the probable error of this difference calculated in the manner explained in preceding sections. Such a quotient gives a convenient index of discrimination. Indexes less than 2 or 3 show that the discriminative capacity of the test for such variables is not significant.

TABLE 21.- DISCRIMINATIVE CAPACITY OF THE INDEXES AS SHOWN BY INTER-MEAN DIFFERENCES IN GRADE 7 AND I HIGH B

Scale	Inter-Mean Difference and Probable Errors for									
	Score		Attempts		Rights		Wrong		Accuracy	
	Diff.	P.E.	Diff.	P.E.	Diff.	P.E.	Diff.	P.E.	Diff.	P.E.
Otis	11.5	2.7	12.3	2.3	12.4	2.7	0.1	2.0	0.0	...
Terman	18.7	3.2	20.7	2.4	16.9	2.6	3.8	1.8	0.01	0.01
Chicago	1.4	1.5	4.4	0.8	1.6	0.9	2.8	0.7	-0.04	0.014
Ave. $\frac{D}{P.E.}$	3.7		6.5		4.3		2.1		1.3	

The five variables in order of their capacity are A,R,S,W, and $\frac{R}{W}$. Thus the groups studied show the greatest difference with respect to speed and the least with respect to accuracy. This result is quite in agreement with common teaching experience. Pupils can be easily made to hurry, but it is exceedingly difficult to train them to be accurate.

While A shows the best capacity for inter-group discrimination, it is not superior to the other variables for differentiating individuals. Score and Rights again appear to be superior to the other indexes for individual discrimination, a property which is more important than inter-group differentiation.

Section 13 Practice Effect with Repetition

Form B of the Terman Scale was given to group I High C and Form A of the same test given the following day. Assuming that these two forms are equally difficult a practice effect for each of the variables may be noted as in Table 22. There is a positive difference between the means for each of the variables except W. This last negative difference also means an improvement on second trial, so that the practice effect is indicated on all of the variables. The last column shows the significance of this gain. The indexes R, A, S, and $\frac{R}{A}$, reveal gains that almost certainly cannot be accounted for by chance fluctuations, while the change in W is in harmony with that of the other variables. Errors and Accuracy show changes of less significance for practice effect.

TABLE 22.- MEANS FOR TERMAN FORMS A AND B WITH GROUP I HIGH C

Variables	M_A (second)	M_B (first)	$M_A - M_B$	Diff÷P.E.$_{diff}$
Score	101.74±1.79	84.11±1.96	17.63±2.65	6.7
Attempts	161.00±1.20	148.19±1.40	12.81±1.84	7.0
Rights	102.41±1.37	87.15±1.56	15.26±2.08	7.3
Wrong	58.59±1.14	61.04±1.14	-2.45±1.60	-1.5
Accuracy	0.63±0.01	0.59±0.01	0.04±0.01	4.4

Part II

Analysis of the Index Variables by Component Tests

The authors' plans of scoring given in Table 3 show that 9 of the tests making up the Otis Scale are scored by the formula S R. These nine tests were therefore chosen for analytical study. The stability of correlations for whole scales has been shown in Part I. In the following sections the intercorrelations of the component tests show a high degree of consistency. The coefficients in general are lower than for whole scales but they indicate the same relationships between index variables. It will also be shown that pooling tests increases both the validity and the reliability of the indexes, an effect which may be roughly forecast by certain predictive formulae.

Section 14 Intercorrelations of Variables for the Otis Components

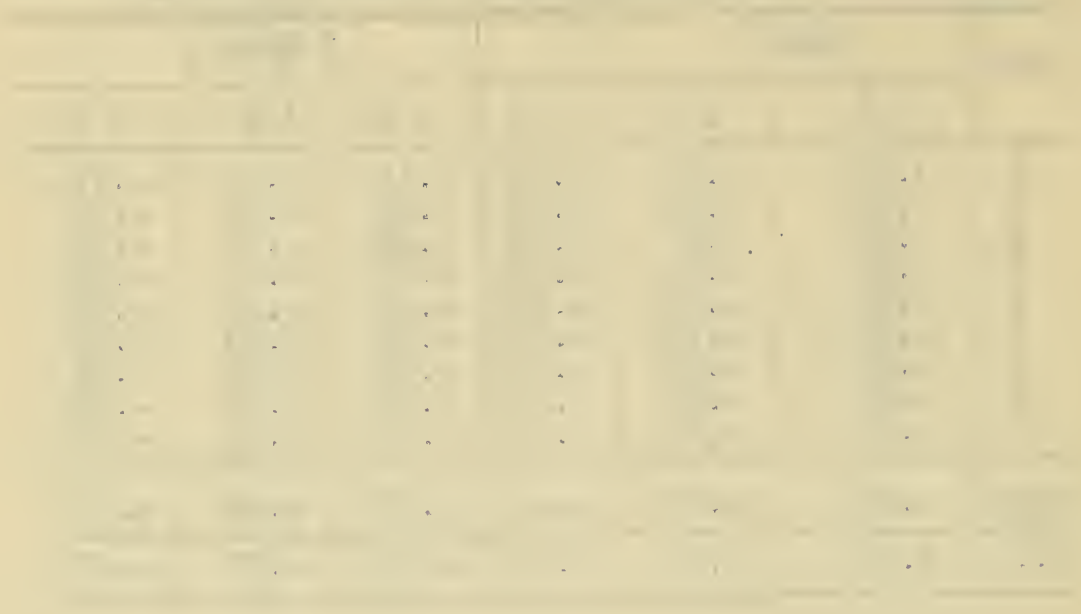
The correlations between Index Variables for the same components are given in Table 23. In the last line of the table the coefficients for all nine tests pooled are given for comparison.

TABLE 23.- CORRELATIONS BETWEEN INDEX VARIABLES ON NINE OF THE COMPONENTS OF THE OTIS SCALE

Test	Grade 7			I High A		
	A×R	A×W	R×W	A×R	A×W	R×W
1	+.37	+.24	-.81	+.19	+.49	-.77
2	+.82	+.43	-.18	+.81	+.55	-.05
4	+.72	+.30	-.42	+.63	+.05	-.75
5	+.72	+.04	-.66	+.70	+.12	-.63
6	+.58	+.48	-.44	+.47	+.48	-.55
7	+.55	+.66	-.28	+.32	+.65	-.52
8	+.28	+.52	-.68	+.36	+.16	-.86
9	+.89	+.20	-.27	+.81	+.18	-.43
10	+.55	+.38	-.54	+.76	+.09	-.85
Mean	+.61	+.36	-.48	+.50	+.33	-.60
All	+.73	+.31	-.51	+.55	+.31	-.70

It is evident that these are higher than the means of the nine correlations on component tests except for Attempts with Errors, in which case the pool gives the lower value. In certain cases, therefore, pooling or lengthening the tests has the effect of increasing the correlation between the indexes. The exception in this instance is worthy of note as a warning against applying general rules for the correlation on lengthened tests. The high degree of consistency in the coefficients indicates that pooling of such components is a justifiable procedure inasmuch as the test material is fairly homogeneous for purposes of indexing.

Intercorrelations between Rights on the nine component parts of the otis scale are given in Tables 24 and 25 for Grade 7 and I High A respectively. Both groups consisted of 50 pupils. All coefficients larger than three times



their probable errors are printed in heavy type.

TABLE 24.- CORRELATIONS BETWEEN RIGHTS ON THE NINE OTIS COMPONENTS FOR GRADE 7

Test	1	2	4	5	6	7	8	9	10
1		+.54	+.40	+.47	+.50	+.34	+.33	+.50	+.24
2	+.54		+.27	+.52	+.39	+.49	+.26	+.27	+.27
4	+.40	+.37		+.35	+.43	+.37	+.40	+.38	+.19
5	+.47	+.52	+.35		+.51	+.38	+.35	+.14	+.18
6	+.50	+.39	+.43	+.51		+.56	+.32	+.36	+.15
7	+.34	+.49	+.37	+.38	+.56		+.13	+.52	+.35
8	+.33	+.26	+.40	+.35	+.32	+.13		+.19	+.29
9	+.50	+.27	+.38	+.14	+.36	+.52	+.19		+.38
10	+.24	+.27	+.19	+.18	+.15	+.35	+.29	+.38	
Mean	+.41	+.38	+.34	+.35	+.40	+.37	+.29	+.34	+.26

A simple calculation will show that this includes all coefficients numerically greater than .27. In Table 24 only 8 of the 36 correlations are not significant, while in Table 25 the same number occur. Most of these low coefficients are found in the correlations with test 10, the mean value for which is lower than for any other test. This one component then appears to be out of harmony with the rest; i.e. to fail to measure the same thing as the other tests of the battery. Inspection of the Otis Scale shows that test 10 is for memory, a trait quite different from those involved in the other components. Except for this one test a fair degree of consistency is found for the coefficients in both tables. The means for all 36 coefficients in each table are +.35 and +.36 respectively.

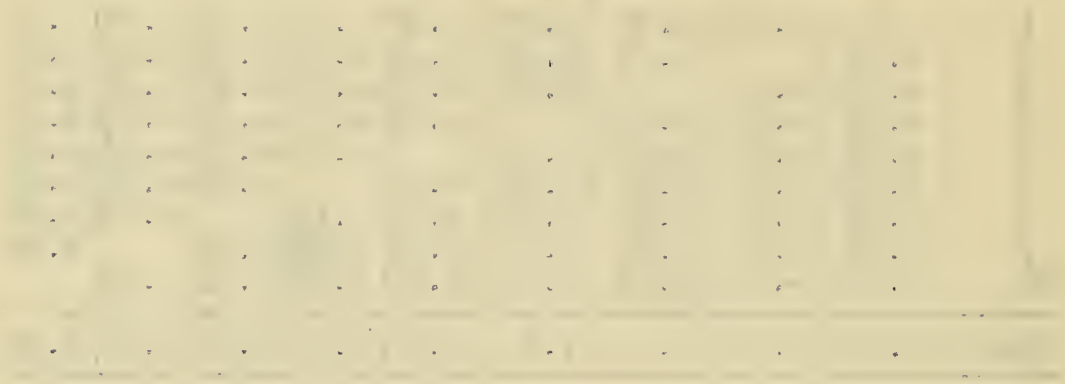


TABLE 25.- CORRELATIONS BETWEEN RIGHTS ON THE NINE OTIS COMPONENTS FOR I HIGH A

Test	1	2	4	5	6	7	8	9	10
1		+.36	+.23	+.45	+.59	+.38	+.5[illegible]	+.42	+.34
2	+.36		+.41	+.36	+.49	+.35	+.56	+.47	+.21
4	+.23	+.41		+.23	+.34	+.38	+.34	+.38	+.32
5	+.45	+.36	+.23		+.53	+.38	+.44	+.17	+.33
6	+.59	+.49	+.24	+.53		+.48	+.52	+.45	+.17
7	+.38	+.35	+.38	+.38	+.48		+.44	+.36	+.02
8	+.5[illegible]	+.56	+.34	+.44	+.52	+.44		+.39	+.19
9	+.42	+.47	+.38	+.17	+.45	+.36	+.39		+.13
10	+.34	+.21	+.32	+.33	+.17	+.02	+.19	+.13	
Mean	+.41	+.4[illegible]	+.33	+.37	+.43	+.35	+.42	+.35	+.21

Tables 26 and 27 show the correlations between errors for the component tests. Only 14 of the 36 coefficients in Table 26 are significant, the mean for the whole table being +.24. Test 10 shows next to the lowest average correlation with the other tests, so that it is of little significance indexed by the rights or wrongs. In Table 27 the coefficients are somewhat higher, the mean of the whole table being +.39. Eight of the 36 correlations are significant, with five of the lowest values appearing with Test 10. It is difficult to explain the difference in correlation for the two groups when indexed by W. Tables 24 and 25 showed mean values nearly identical, but the difference between the mean coefficients for W is too large to be ascribed to chance. One explanation of this difference may be found in the fact that Group I High A made more errors than Grade 7 (See Table). The effect of this was to give less jamming in the contingency tables with a resultant higher correlation.

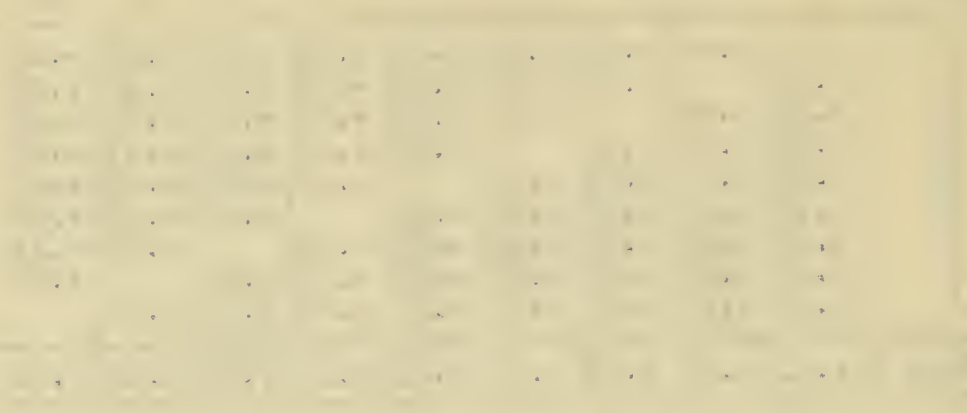


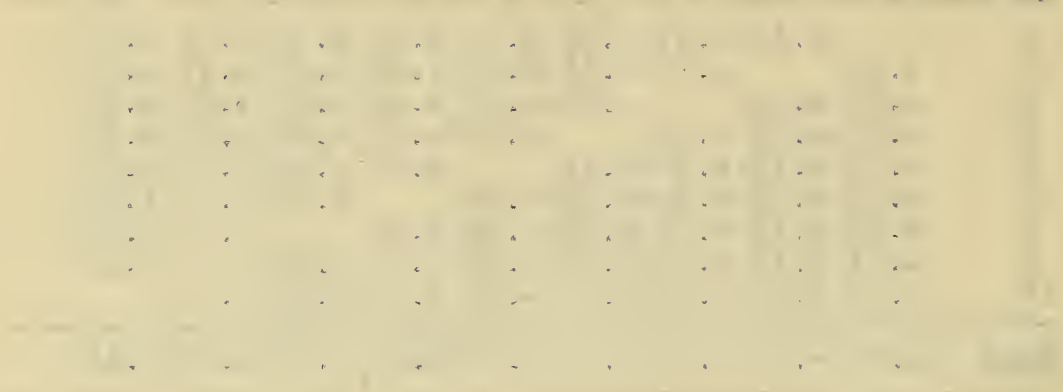
TABLE 26.- CORRELATIONS BETWEEN ERRORS ON THE NINE OTIS COMPONENTS FOR GRADE 7

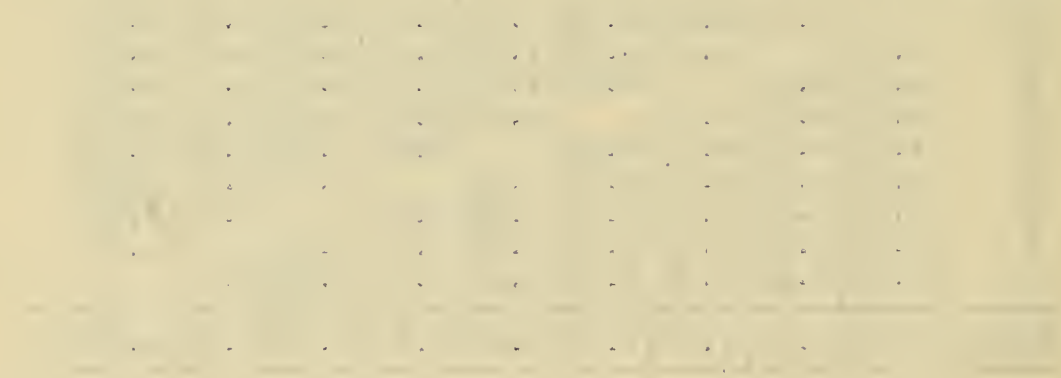
Test	1	2	4	5	6	7	8	9	10
1		+.16	+.17	+.49	+.11	+.29	+.30	-.01	+.15
2	+.16		+.24	+.15	+.18	+.35	+.16	+.17	+.13
4	+.17	+.24		+.32	+.30	+.44	+.37	+.33	+.28
5	+.49	+.15	+.32		+.18	+.45	+.35	+.09	+.19
6	+.11	+.18	+.30	+.18		+.40	+.24	+.18	+.22
7	+.29	+.35	+.44	+.45	+.40		+.43	+.21	+.23
8	+.30	+.16	+.37	+.35	+.24	+.43		+.21	+.14
9	-.01	+.17	+.33	+.09	+.18	+.21	+.21		+.12
10	+.15	+.13	+.28	+.19	+.32	+.23	+.14	+.12	
Mean	+.21	+.19	+.31	+.28	+.23	+.35	+.27	+.16	+.18

TABLE 27.- CORRELATIONS BETWEEN ERRORS ON THE NINE OTIS COMPONENTS FOR I HIGH A

Test	1	2	4	5	6	7	8	9	10
1		+.44	+.42	+.28	+.38	+.58	+.49	+.46	+.39
2	+.44		+.44	+.36	+.54	+.57	+.36	+.51	+.18
4	+.42	+.44		+.30	+.37	+.54	+.51	+.39	+.22
5	+.28	+.36	+.30		+.48	+.52	+.46	+.12	+.29
6	+.38	+.54	+.37	+.48		+.48	+.49	+.26	+.16
7	+.58	+.57	+.54	+.52	+.48		+.54	+.50	+.31
8	+.49	+.36	+.51	+.46	+.49	+.54		+.21	+.24
9	+.46	+.51	+.39	+.12	+.26	+.50	+.21		+.10
10	+.39	+.18	+.22	+.29	+.16	+.31	+.24	+.10	
Mean	+.48	+.42	+.40	+.35	+.40	+.51	+.41	+.32	+.24

Comparison of the four tables above shows that the intercorrelation of errors on the 9 component tests is about as high as for Rights. It may be noted also that none of the correlations are as high as .6 while the means in all the tables are less than .4. Such coefficients are not considered high. The correlations corresponding for whole scales as given in Table 6 are +.74 for Rights and +.69 for Errors. Clearly then the cumulating of tests to



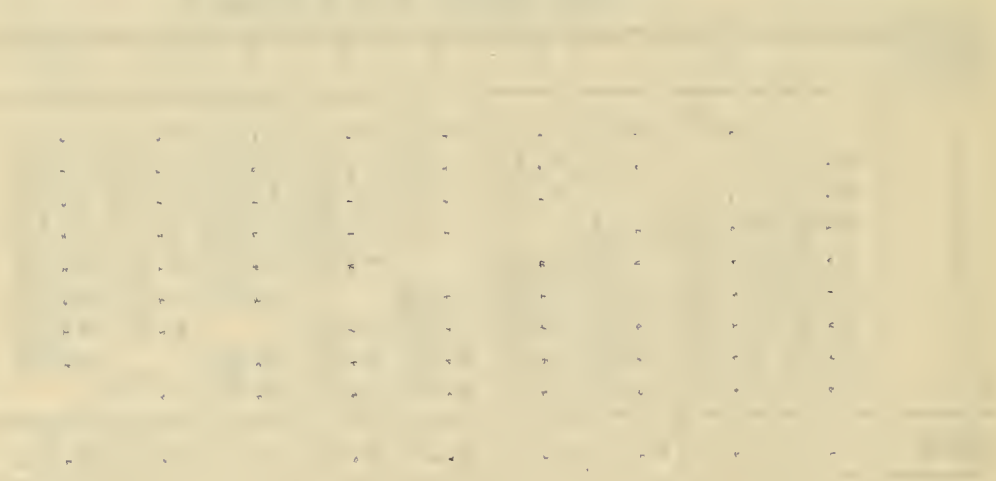


form what has been called a scale score has the effect of raising the correlation or, in other words, lengthening a test increases its reliability for these indexes. A more detailed discussion of this point will appear later.

Correlations between attempts on the nine components have been worked out for one group and are given in Table 28. Of the 36 coefficients, 27 are significant, the lowest average again occurring for Test 10 with each of the others. By three modes of indexing, then, this test shows up as distinct in type from the rest. The mean correlation for the whole table is +.32 which may be compared with the mean coefficient of .56 in Table 6. Lengthening the test also increases correlation when Attempts are employed as the index variable. The comparison is only a rough one, however, for somewhat different scales and groups are employed in the two cases.

TABLE 28.- CORRELATIONS BETWEEN ATTEMPTS ON THE NINE OTIS COMPONENTS FOR I HIGH A

Test	1	2	4	5	6	7	8	9	10
1		+.30	+.41	+.48	+.28	+.35	+.40	+.21	+.31
2	+.30		+.32	+.13	+.38	+.41	+.39	+.29	+.26
4	+.41	+.32		+.41	+.29	+.29	+.32	+.35	+.20
5	+.48	+.13	+.41		+.31	+.29	+.34	+.37	+.18
6	+.28	+.38	+.29	+.31		+.52	+.27	+.40	+.26
7	+.35	+.41	+.29	+.29	+.52		+.37	+.22	+.33
8	+.40	+.39	+.32	+.34	+.27	+.37		+.40	+.25
9	+.21	+.29	+.35	+.37	+.40	+.22	+.40		+.08
10	+.31	+.26	+.20	+.18	+.26	+.33	+.25	+.08	
Mean	+.34	+.31	+.32	+.31	+.34	+.35	+.34	+.29	+.24



In Part I, Table 5, R and W showed a mean correlation of -.49 for whole scales. The corresponding coefficients for the Otis components are given in Table 29 with a mean of -.21. The increase in correlation by pooling is again evident. The correlations in the principal diagonal are between Rights and Errors on the same test and are therefore larger than the rest, the mean being -.48. The remainder of the table gives correlations for all possible combinations of Rights and Wrongs on the nine tests two at a time. All but three of the 81 coefficients are negative while just one third of them are significant according to the usual rule.

TABLE 29.- CORRELATIONS BETWEEN RIGHTS AND ERRORS ON THE NINE OTIS COMPONENTS FOR GRADE 7 AND I HIGH A

Wrongs	Rights									Mean
	1	2	4	5	6	7	8	9	10	
1	-.81	-.3[illegible]	-.[illegible]4	-.24	-.33	-.41	-.27	-.28	-.31	-.35
2	-.[illegible]1	-.18	-.17	-.[illegible]6	-.15	-.[illegible]	-.1[illegible]	-.1[illegible]	-.25	-.17
4	-.11	-.11	-.42	-.[illegible]9	-.11	-.05	-.18	-.09	-.02	-.13
5	-.44	-.27	-.4[illegible]	-.66	-.44	-.29	-.51	-.07	-.18	-.36
6	-.11	-.[illegible]8	-.4[illegible]	-.[illegible]9	-.44	-.24	-.11	-.2[illegible]	-.11	-.18
7	-.2[illegible]	-.1[illegible]	-.41	-.[illegible]3	-.[illegible]1	-.2[illegible]	-.[illegible]1	-.[illegible]9	-.13	-.23
8	-.[illegible]4	-.2[illegible]	-.42	-.[illegible]8	-.[illegible]3	-.17	-.[illegible]3	-.07	-.[illegible]6	-.28
9	-.[illegible]4	-.[illegible]3	-.31	+.[illegible]	-.1[illegible]	-.[illegible]	-.1[illegible]	-.27	-.[illegible]1	-.[illegible]9
10	+.[illegible]7	+.[illegible]4	-.[illegible]9	-.07	-.15	-.17	-.13	-.[illegible]8	-.54	-.13
Mean	-.[illegible]5	-.1[illegible]	-.31	-.[illegible]2	-.[illegible]5	-.1[illegible]	-.[illegible]6	-.14	-.1[illegible]	-.21

This suggests that the criterion of three times the probable error is too stringent for tests of this kind. If twice the probable error were adopted in the present case all coefficients over.18 would be significant including

five ninths of the total number, while all of those greater than one probable error or over .10 will include 56 out of 81. For the last case 25 coefficients are less than one probable error, yet 22 of them are negative in sign. Thus a coefficient less than one probable error appears to give assurance of negative correlation beyond the expectation from the usual rule of even chance; i.e. the probability of significance from the data appears to be greater than by theory. In any case, highly consistent negative correlation is exhibited by the whole array.

Section 15 Correlations of the Otis Components with Age

The correlations of the age factor with each of the Otis components are similar to those for whole scales. Table 30 shows higher correlations for the nine tests pooled than for the mean of the tests. Here again the effect of adding tests is to increase correlation. The formula for estimating the correlation of the pool of "n" tests with a criterion may be written in the form:

$$r_{(x_1+x_2+\cdots x_n)c} = \frac{r_{x_1c}+r_{x_2c}+\cdots\quad r_{x_nc}}{\sqrt{n+2(r_{x_1x_2}+r_{x_1x_3}+\text{ to }\frac{(n)(n-1)}{2}\text{ terms})}} \quad \text{(Theorem 4 Appendix)}$$

Considering $x_1 x_2, \cdots x_{10}$ as Rights for Otis Grade 7 and age as a criterion, the constants in Tables 24 and 30 give for this coefficient, $r_{(R_1+R_2+\cdots R_{10})\,age} = -.39$

a value identical with that obtained by pooling the nine components.

TABLE 30.- CORRELATIONS BETWEEN AGE AND THE INDEX VARIABLES ON THE OTIS SCALE FOR GRADE 7 AND I HIGH A

Test	Grade 7			I High A		
	Age x A	Age x R	Age x W	Age x A	Age x R	Age x W
1	-.43	-.30	+.13	-.38	-.23	+.02
2	-.29	-.43	+.18	-.[illegible]	+.31	+.28
4	-.3[illegible]	-.24	-.07	-.19	-.29	+.21
5	-.33	-.42	+.25	-.39	-.47	+.23
6	-.23	-.33	+.25	-.21	-.52	+.33
7	-.[illegible]	-.26	+.21	-.[illegible]	-.26	+.19
8	-.[illegible]	-.26	+.06	-.19	-.41	+.35
9	-.1[illegible]	-.19	+.[illegible]	-.33	-.44	+.34
10	-.3[illegible]	+.03	-.35	-.[illegible]	+.01	-.04
Mean	-.25	-.26	+.25	-.23	-.3[illegible]	+.2[illegible]
All	-.38	-.39	+.27	-.[illegible]	-.32	+.34

Similar coefficients are given in Table 31. The differences between predicted and actual values are in no case significant. The above formula, then, appears to be a useful one in predicting the validity (correlation with a criterion) of tests by pooling components. It is to be noted also that the formula will have high values for large values of r_{xc} and small values of r_{xx} .

TABLE 31.- PREDICTED AND ACTUAL CORRELATIONS BETWEEN AGE AND INDEX VARIABLE

Group	Variables	Predicted Value	Actual Value
Grade 7	Rights Age	-.39	-.38
I High A	Rights Age	-.4[illegible]	-.52
Grade 7	Wrongs Age	+.[illegible]	+.27
I High A	Wrongs Age	+.3[illegible]	+.34
I High A	Attempts Age	-.35	-.32

To obtain a scale of high validity, therefore, component tests should be selected which have high correlation with the criterion but low correlation among themselves. Thorndike[a] justified the use of tests with low inter- correlations on the ground that they are repetitive; i.e. measures of the same fact. The above formula, however, will give high validity because the inter-test correlations occur in the denominator, and low values will thus raise the value of the fraction. The basis of test selection, then, appears to be mathematical, rather than psychological. Fortunately, however, the two bases agree.

Section 16 The Application of Reliability Formulae to Component Tests

In the preceding section it was shown that pooling component tests has the effect of increasing the validity of the total scale; i.e. to the extent to which it correlates with a criterion. The pooling of tests will next be shown to have a similar effect upon the reliability of a scale; i.e. the correlation between two forms of the same test.

It will be recalled that two forms of the Terman Scale were given to Group I High C on successive days. Reliability coefficients have been calculated for each of the 10 component tests and for all combined to give the total score. The results are given in Table 32.

---

[a] Memoirs of the National Academy of Sciences, Vol.XV, p. 316

TABLE 32.- RELIABILITY COEFFICIENTS FOR THE TERMAN SCALE BY COMPONENTS AND TOTAL SCORE

Test	Formula	Correlation between Forms A and B	Rank
1	R	+.630	7
2	2R	+.809	3
3	R-W	+.682	6
4	R	+.900	1
5	2R	+.852	2
6	R-W	+.482	10
7	R	+.683	5
8	R-W	+.530	8
9	R	+.514	9
10	2R	+.792	4
Mean		+.679	
All		+.916	

Certain of the individual tests reveal a high degree of reliability, especially test 4 (logical selection) with a coefficient of .900. It will also be observed that some of the lowest correlations occur with tests scored R-W. This point will be dealt with more fully in a following section. The mean of the reliability coefficients on the tests is +.679, while the correlation for total Score on the two forms is +.916, so that pooling the tests has the effect of increasing reliability.

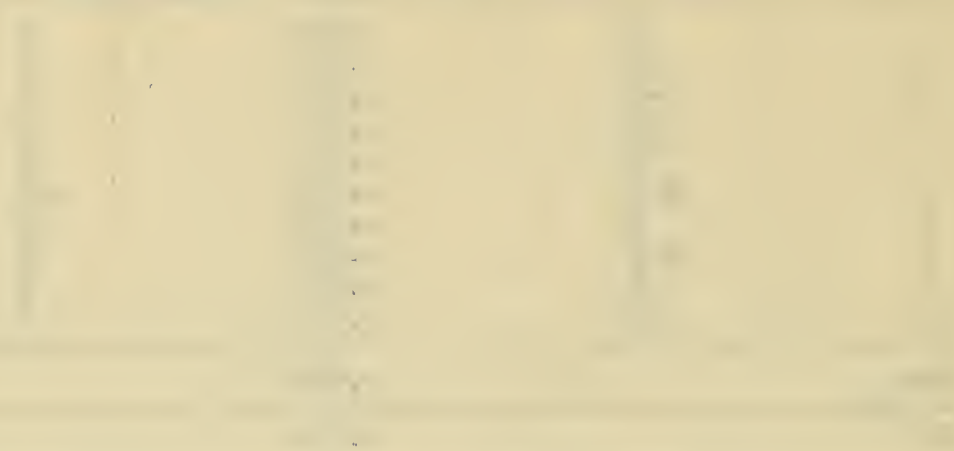
A predictive formula given by Brown[a], and also implied in Spearman's General Theorem[b] may be given in the form,

$$r_{NN} = \frac{N r_{11}}{1+(N-1) r_{11}} \qquad \text{(Theorem 5, Appendix)}$$

where r_{11} is the correlation between two tests or the aver-

a. William Brown, Essentials of Mental Measurement, Cambridge University Press, London, 1911

b. C. Spearman, Correlation of Sums and Differences, Brit. Jour. of Psy., Vol.5, 419-426



age of several, and N the number of tests thus amalgamated. In the present example the average correlation from the first three tests is. 71, In order to predict the reliability coefficient for 10 such tests, it is only necessary to substitute these values in the above formula giving,

$$r_{(10)(10)} = +.96$$

The value from actual amalgamation is .92. Similarly, a calculation based upon the average of all 10 tests also gives

$$r_{(10)(10)} = +.96$$

The use of the formula in these cases, then, gives considerable over prediction.

In order to test the applicability of Brown's Formula more fully and to analyze more fully the effect of pooling tests on reliability, a more detailed procedure is next employed. Reliability coefficients on cumulated tests are obtained in two ways: The scores for tests 1 and 2 on each form of the Terman Scale are added, and the correlation determined; next tests 1,2,and 3 are pooled and the two forms correlated, and so on until all 10 tests have been cumulated in this fashion. The second procedure is to begin with tests 10 and 9 and amalgamate in the reverse direction. These empirical results are then compared with theoretical values obtained by substituting r = .68 and N from 1 to 10 in Brown's Formula. Table 33 gives the results of this lengthy calculation.

TABLE 33.- THEORETICAL AND ACTUAL RELIABILITY COEFFICIENTS OBTAINED FROM BROWN'S FORMULA AND BY SUCCESSIVE CUMULATION OF THE TEN THURIAN COMPONENTS

Number of Tests Cumulated	Theoretical Value	Order of Cumulation	
		1 to 10	10 to 1
1	+.68	+.64	+.70
2	+.82	+.81	+.79
3	+.87	+.87	+.83
4	+.90	+.91	+.86
5	+.92	+.90	+.84
6	+.93	+.88	+.86
7	+.94	+.89	+.87
8	+.94	+.87	+.87
9	+.95	+.91	+.90
10	+.96	+.92	+.92

Inspection of the table shows a rough agreement in the three series. Figure 5 which is based on the table brings out the comparisons more clearly. The three curves show a rapid initial rise up to 4 cumulated tests and then a more gradual increase to the maximum value. The more rapid rise of the curve cumulated from tests 1 to 10 than the reverse one, is no doubt due to the greater reliability of the first few tests as indicated in Table 32.

While the general agreement in the three curves is evident, nevertheless there is a very clear tendency for the theoretical curve calculated for r = .68 to give an over prediction beyond 4 or 5 cumulated tests. This may be partly due to the unequal values of the individual reliability coefficients given in Table 32, but whatever the cause, the use of Brown's Formula for prediction in a case of this kind is open to question. The equation

To follow - p 57

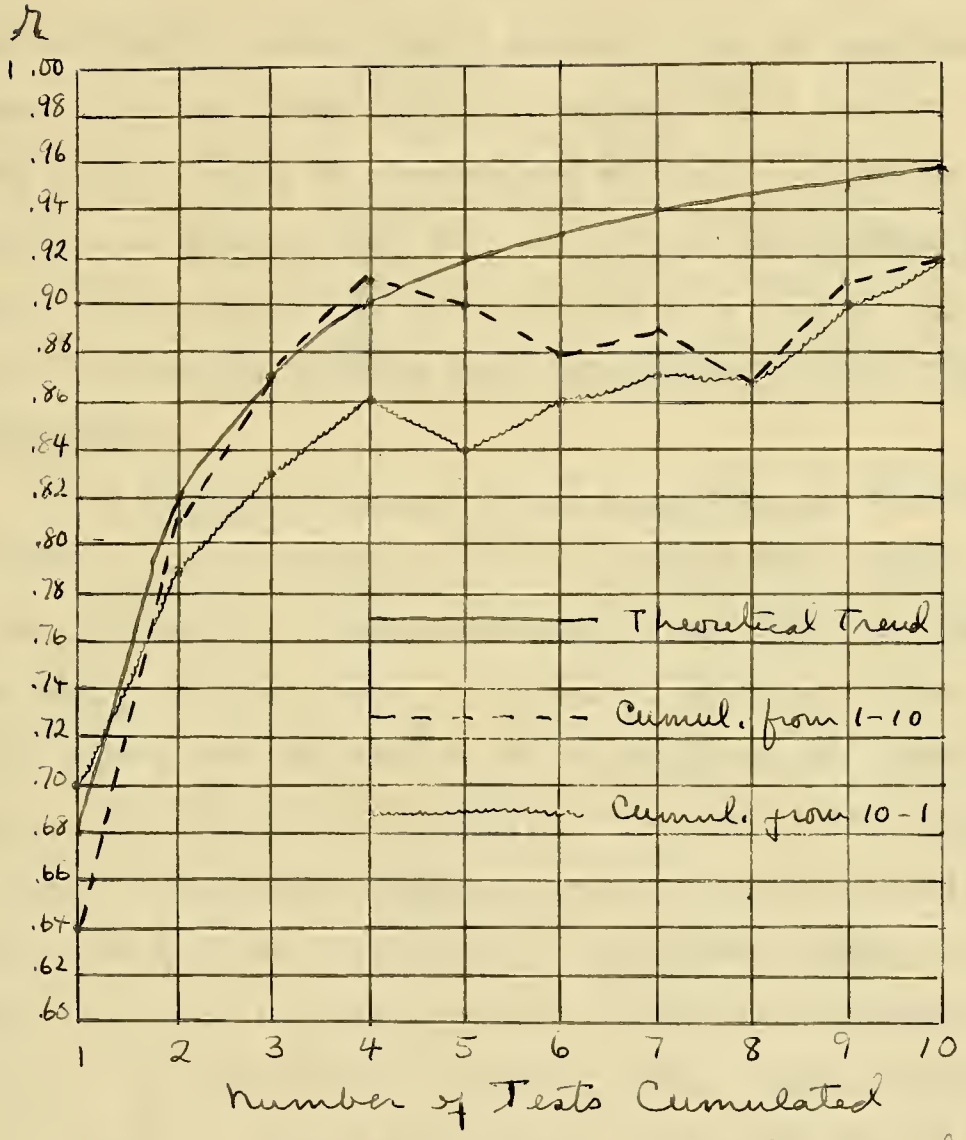


Fig 5. Theoretical and Actual Reliability Trends based on Ten Terman Components.

and corresponding theoretical curve indicate that to get any desired degree of reliability with +1.00 as an upper limit, it is only necessary to amalgamate tests indefinitely. This is, of course, absurd. The formula gives an over prediction fairly early in the series of cumulated tests. From the above tables it appears that four or five typical tests of the battery will give almost as reliable an index as the pool of all ten components. This result would account in part for the high reliability of such tests as tha Chicago Scale consisting of only five components.

This problem is one of great importance in test construction. If intelligence can be indexed with almost as great accuracy by a short scale as by one twice as long, the saving in time alone is enormous. Moreover if the short series can be shown to be as valid as the longer one by correlation with criteria, the abbreviated method is further justified. Inasmuch as no suitable criterion other than age was available for the present data, the check cannot be rigidly applied. The age correlations by half and by whole scales, however, agree almost exactly (-.37, -.39), so that with age as a criterion, the five test battery is as valuable as the ten test scale.

Section 17 Summary of Analysis of Components

The relationships found between index variables in Part I are verified for component tests. These coefficients are in general lower than by whole tests, so that pooling has the effect of increasing the correlation between indexes. Inter-correlations between components for R,W, and A reveal a high degree of consistency for such short tests but are less stable than for similar coefficients by whole tests. Furthermore the consistent batteries of correlations even for R and W on different tests indicate a high degree of homogeneity in the test material with the possible exception of Test 10.

In addition to raising the correlation between index variables, pooling tests also has the general effect of increasing the validity and reliability of a scale within certain limits. Predictive formulae are useful in this connection but are likely to give an over-estimate of the correlation to be expected by pooling. Moreover the physical endurance of the children determines the maximum length of the tests at a sitting, so that the formulae are limited in application. The gain in validity and reliability is rapid on pooling the first few tests, but the point is soon reached where the addition of similar material affects the correlations but slightly. The results indicate that a battery of four or five carefully selected components will give an index with substantially the same reliability as a scale twice that length.

Part III Scoring Formulae

Section 18 The Linear Form, S = a (R + bW)

a. Formulae with Highest Validity

In Parts I and II it has been shown that the scoring formulae employed by the authors of the scales have little effect upon the resultant scores when a number of components are pooled. The Terman Scale with the components scored by the three formulae, S = R, S = 2R, and S = R-W has a correlation of +.96 with the score obtained by using S-R on all ten components (Tables 4 and 5). The amalgamated score then, is not very sensitive to such changes in the component scoring formulae and simple forms are recommended on these grounds. The single component, however, is much more violently affected by changes in the formulae employed to index it. Changes in weights which affect the pooled score but slightly, will be found to have a pronounced effect upon the individual components.

Table 4 which gives the various component scoring formulae used by the authors of the scales, includes only formulae of the linear type; i.e. equations of the first degree in the variables employed. These variables are R and W, so that the most general formula used may be written,

(1) $S = a(R + bW) = aR + cW$

where a, b, and c, are constants. It was also noted in Section 3 that the relationship, $A = R + W$

makes it possible to express this formula in terms of R and A or W and A. Formula (1), however, has been so generally employed that it will be adopted here for further analysis. Formulae expressed in terms of the other variables may be obtained by substitution if they are required.

The question immediately arises as to the best values to assign the constants a and c in equation1. A general solution of this problem may be obtained by the method of least squares. Values for R and W are obtained for each of the N individuals of a given population. Assuming that a criterion, K, is the best measure of such determinations, a set of N equations may be formed,

$$K_1 = a_1 R_1 + c_1 W_1$$
$$K_2 = a_2 R_2 + c_2 W_2$$
$$- - - - - - - - -$$
$$K_N = a_N R_N + c_N W_N$$

where the K's, R's, and W's are known, and a's and c's are to be determined so as to minimize the inconsistency in the equations which is assumed to be due to imperfect measurement.

Next $V_1, V_2, - - - V_N$ will be written for the differences between $K_1, K_2 - - K_N$ and the values obtained from the best determinations for the a's and c's:

$$a_1 R_1 + c_1 W_1 - K_1 = V_1$$
$$a_2 R_2 + c_2 W_2 - K_2 = V_2$$
$$- - - - - - - - - - -$$
$$a_N R_N + c_N W_N - K_N = V_N$$

These differences or "residuals" are assumed to be

normally distributed. While the assumption is open to question for data of this type, it is nevertheless the best that can be made. The most probable values for a and c next require that the sum

$$V_1^2 + V_2^2 + \cdots V_N^2 = \text{a minimum.}$$

The remainder of the procedure consists in setting up the "normal equations" in the usual way. Transferring the variables to their respective means, and setting up these equations gives,

$$a\Sigma\lambda^2 + c\Sigma\lambda w = \Sigma k\lambda$$

$$a\Sigma\lambda w + c\Sigma w^2 = \Sigma kw$$

Since, $\sigma_x^2 = \frac{\Sigma x^2}{N}$, and $r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y}$, these equations may be written in the form,

$$a\sigma_\lambda + c\, r_{\lambda w}\sigma_w = r_{k\lambda}\sigma_k$$

$$a\, r_{\lambda w}\sigma_\lambda + c\,\sigma_w = r_{kw}\sigma_k$$

Solving these equations for a and c gives,

$$(3) \qquad a = \frac{\sigma_k\left(r_{kw}r_{\lambda w} - r_{k\lambda}\right)}{\sigma_\lambda\left(r_{\lambda w}^2 - 1\right)}$$

$$(4) \qquad c = \frac{\sigma_k\left(r_{k\lambda}r_{\lambda w} - r_{kw}\right)}{\sigma_w\left(r_{\lambda w}^2 - 1\right)}$$

The value $\frac{c}{a}$ may then be written

$$(5) \qquad C = \frac{c}{a} = \frac{\sigma_\lambda\left(r_{k\lambda}r_{\lambda w} - r_{kw}\right)}{\sigma_w\left(r_{kw}r_{\lambda w} - r_{k\lambda}\right)}$$

This last result has been obtained by Thurstone[a] as a

a. L.L. Thurstone, A Scoring Method for Mental Tests, Psy. Bull., Vol. XVI, No.7, July, 1919.

value for C in the formula $S = R + CW$ such that the correlation r_{KS} is a maximum; i.e. C or $\frac{c}{a}$ is determined in such a way as to possess the highest validity with a criterion. The formula for this correlation is,

$$(6) \qquad r_{KS} = \frac{r_{Kr}\sigma_r + C\, r_{Kw}\sigma_w}{\sqrt{\sigma_r^2 + 2C\, r_{rw}\sigma_r\sigma_w + C^2\sigma_w^2}}$$

Thurstone also makes use of Yule's[a] equation for multiple correlation to obtain an expression for the highest correlation with the linear formula $S = R + CW$. This result may be written

$$(7) \qquad R_{K(ar+cw)} = \sqrt{\frac{r_{Kr}^2 + r_{Kw}^2 - 2\, r_{Kw} r_{Kr} r_{rw}}{1 - r_{rw}^2}}$$

The actual procedure involved in determining the constants a and c will then be as follows:

1. Give the test to a group and obtain also the criterion of validity against which the formula is to be checked.
2. Score the test for R and W and compute the constants r_{Kr}, r_{Kw}, r_{rw}, σ_r, and σ_w (σ_K is not required if Formula (5) is employed).
3. Substitute these last results in equations (3),(4), and (5) and obtain the formulae $S = aR + cW$ or $S = R + \frac{c}{a}W$ (differing only by factor of proportionality, $\frac{1}{a}$).
4. To predict the highest correlation obtainable with these formulae, substitute the computed constants in equation (7).

---

a. G.U.Yule, Introduction to Statistics, C.Griffin, London 1919, 248.

b. Limitations in the Use of the Formula $S = aR + cW$.

In section 3, two plans for administering tests were described. According to the first, the time is fixed and Attempts, Rights, and Wrongs allowed to vary; according to the second, the number of Attempts is fixed, while Time, Rights, and Wrongs are recorded. Two of the index variables are thus alternately controlled by the method of administering the test.

The formula $S = aR + cW$, with constants determined by highest validity with the criterion, serves very well for tests given according to the plan of fixing the time. All of the tests employed in the three intelligence scales are of this type and hence no difficulty is encountered.

For tests administered by fixing the Attempts, however, the above formula is inadequate. This arises from the fact that S is no longer a function of two independent variables, but of only one. If Attempts are constant and $A = R + W$, then $W = -R + d$ where d is constant. Substituting the value for W in the expression $S = aR + cW$ gives $S = aR + c(d - R)$ or $S = (a - c)R + cd$. Thus the scoring formula is independent of , and no matter what value is assigned to c (with the exception of $c = a$, for which value the correlation is zero) the correlation r_{KS} given by formula (6) is equal to r_{KR}. This last result follows from the fact that $r_{x(ay+b)} = r_{xy}$ where a and b are constants (Theorem 1, Appendix)

In other words if Attempts are constant, the scoring formula $S = a R$ has the same validity with a criterion as any linear function of R and W (with the exception $R + W$ for which $r_{ks} = 0$).

A further limitation in the use of the formula $S = R + CW$ lies in its sensitiveness. The value of C as determined by Formula (5) depends upon σ_r and σ_w which in turn depend upon the proportion of Rights and Wrongs in the group. Thus while the value for C is the best prediction for the particular group, the value will very likely differ very materially in other groups with differing percentages of Wrongs. This means that the formula $S = R + C.W$ may be used with safety only (for tests administered with time fixed) when the value for C has been determined after the test has been given to the particular group, not before. Thurstone[a] discusses the sensitiveness of the formula, but is unwilling to admit its limitations. Table 34 which is based on extremely lengthy and careful calculation, indicates wide variation in the determinations of C for the 9 Otis components on Grade 7 and I High A. The average difference in the values of C for the two groups is .549 and in only one test is the difference less than 1. These results point clearly to the conclusion that the above formula has little general merit; i.e. the value for C from one group cannot be safely assumed to hold for another.

---

[a] L.L.Thurstone, loc cit.

TABLE 34.- VALUES FOR C IN THE FORMULA S R+CW DETERMINED FOR GROUPS GRADE 7 AND I HIGH A ON THE NINE COMPONENTS OF THE OTIS SCALE (AGE CRITERION)

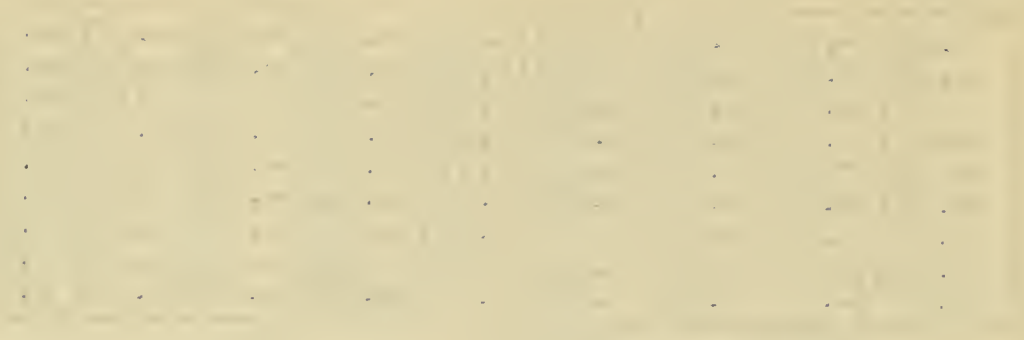
Test	Grade 7						I High A		
	Age×R	R×W	Age×W	σ_R	σ_W	C_1	Age×R	R×W	Age×W
1	-.381	-.808	+.126	2.621	2.506	+.507	-.227	-.767	+.017
2	-.423	-.176	+.182	2.813	1.794	-.414	-.308	-.047	+.282
4	-.237	-.122	-.066	4.190	2.253	+.091	-.220	-.752	+.212
5	-.417	-.662	+.248	3.477	2.409	+.169	-.468	-.627	+.230
6	-.378	-.436	+.042	2.742	2.076	+.054	-.523	-.552	+.325
7	-.261	-.276	+.214	2.758	2.059	-.034	-.256	-.518	+.189
8	-.359	-.675	+.072	3.300	2.597	+.413	-.408	-.857	+.327
9	-.123	-.274	+.000	5.116	2.407	+.570	-.443	-.469	+.244
10	+.038	-.540	-.347	2.982	2.570	2.376	+.006	-.051	-.040

TABLE 34 CONTINUED

Test	I High A			Diff.
	σ_R	σ_W	C_2	$C_1 - C_2$
1	1.918	2.160	+0.052	-.455
2	2.044	2.003	-1.289	+.875
4	4.605	3.605	+0.059	+.032
5	3.072	2.427	+0.272	-.112
6	2.514	2.526	-0.105	+.409
7	2.871	3.578	-0.287	-.347
8	3.830	2.665	+0.186	+.227
9	4.599	3.715	-0.270	+.248
10	3.056	3.079	+1.236	+1.149

c. The Use of the Formula $S = R + CW$

In test material where two alternatives are given for each item and guessing therefore possible the formula $S = R - W$ (i.e. $C = -1$) has been frequently adopted. Similar formulae are used when the number of choices is greater. These expressions are assumed to correct for guessing element involved. According to the above equation a person guessing blindly on all of the items will get half of them right by chance, and hence a zero score which he deserves.



For actual guessing, then, the formula penalizes justly; but it also penalizes for errors which are not due to guessing, and hence unjustly. As a result, for very difficult material, nearly all of the scores may be negative.

A number of experimental attempts have been made to determine the amount of guessing in tests of this sort. After administering a set of True-False tests of the syllogystic reasoning type , the writer asked the pupils on which items they had guessed. The number guessed was about two per cent of the total number of errors made, and of those items guessed only 8^ per cent were wrong. For such groups and tests, the penalty attached to errors by the formula $R - W$ is enormously too great. It may be noted, however, that the children did not always know whether or not they had guessed on any item. Reasoning and guessing are often indistinguishable, and who has not credited himself with reasonong when he has only made a lucky guess.

Instead of assuming that a penalty should be attached for guessing, Thurstone proposes to use the formula with the value of C to be determined according to validity as above. This method appears to be preferable to that of a priori determination, when tests are administered with the time fixed. If Attempts are fixed, however, the formula becomes independent of W as has just been shown, and all values of C (except $C = +1$) gives the same correlation $r_{K(R+CW)} = r_{KR}$.

This point is of great practical importance because most tests of the True-False type are administered so that all may finish; i.e. Attempts constant. According to the above results, all formulae of the type $S = R - \frac{1}{n} W$ give the same correlation with a criterion as is obtained by using Rights alone; i.e. $S = R$. Inasmuch as the expression $R - \frac{1}{n} W$ does not correct adequately for guessing, and has the same validity as $S = R$ for Attempts constant, the writer believes it should be abandoned in favor of the simpler form.

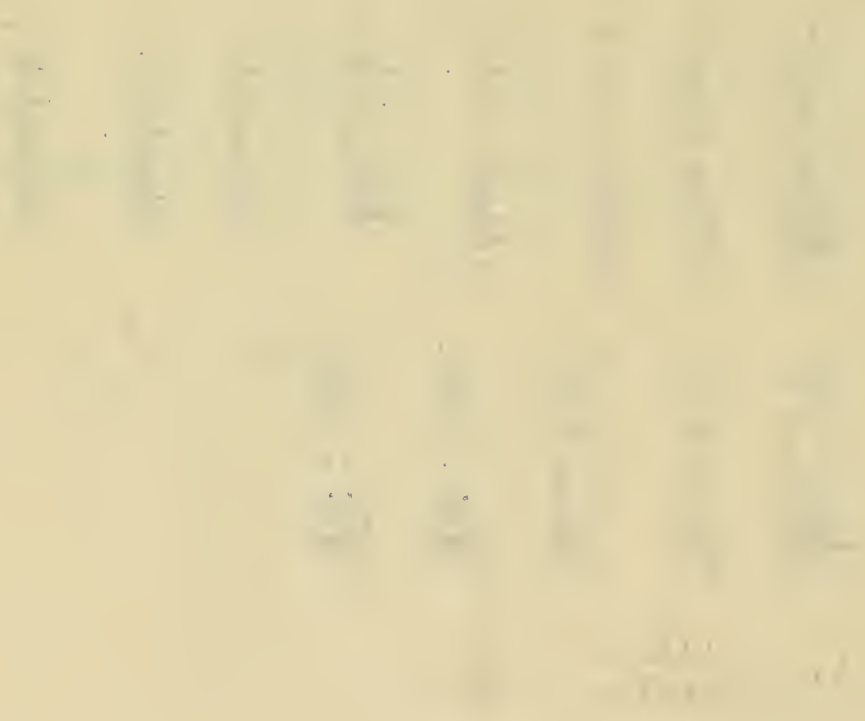
The following example illustrates the foregoing discussion.

K	R	W	R-W	k	n	w	n-w
10	2	8	-6	-20	-2	2	-4
20	2	8	-6	-10	-2	2	-4
30	5	5	0	0	1	-1	2
40	4	6	-2	10	0	0	0
50	7	3	4	20	3	-3	6
150	20	30	-10				
30	4	6	-2				

k^2	n^2	$(n-w)^2$	kn	k(n-w)
400	4	16	40	80
100	4	16	20	40
0	1	4	..	..
100	0	0	..	..
400	9	36	60	120
1000	18	72	120	240

$$\Lambda_{kn} = \frac{120}{\sqrt{18000}} = \frac{2}{\sqrt{5}}$$

$$\Lambda_{k(n-w)} = \frac{240}{\sqrt{72000}} = \frac{2}{\sqrt{5}}$$



Section 19 Simple Ratios

a. The Correlation Between Speed and Accuracy

The formulae $S = \frac{A}{T}$ and $S = \frac{R}{A}$ may be conveniently employed to index Speed and Accuracy respectively. The latter form has been used extensively in the first two parts of this study and has been found to have valuable properties which other indexes do not possess. With Time constant as in the intelligence scales discussed, the formula for Speed reduces to $S = cA$; i.e. the Attempts give the measure of Speed directly when Time is fixed. When Attempts are constant, the Speed is given by the reciprocal of the Time (in suitable units).

The general expression for the correlation between two ratios $\frac{X}{Y}$ and $\frac{Z}{W}$ may be written in the form

$$(3) \qquad \rho_{\frac{X}{Y}\frac{Z}{W}} = \frac{\rho_{XZ}V_XV_Z + \rho_{YW}V_YV_W - \rho_{XW}V_XV_W - \rho_{YZ}V_YV_Z}{\sqrt{\left[V_X^2 - 2\rho_{XY}V_XV_Y + V_Y^2\right]\left[V_Z^2 - 2\rho_{ZW}V_ZV_W + V_W^2\right]}}$$

where the V's are coefficients of variation given by the formula $V = \frac{100\,S.D.}{M}$ (Theorem 2 Appendix). For the two ratios $\frac{A}{T}$ and $\frac{R}{A}$ this expression becomes

$$\rho_{\frac{A}{T}\frac{R}{A}} = \frac{\rho_{AR}V_AV_R + \rho_{TA}V_TV_A - \rho_{AA}V_AV_A - \rho_{RT}V_RV_T}{\sqrt{\left[V_A^2 - 2\rho_{AT}V_AV_T + V_T^2\right]\left[V_R^2 - 2\rho_{RA}V_RV_A + V_A^2\right]}}$$

For T = constant, V_T = constant, and all correlations with T are zero, therefore,

$$\rho_{A\frac{R}{A}} = \frac{\rho_{AR}V_AV_R - V_A^2}{\sqrt{V_A^2\left[V_R^2 - 2\rho_{RA}V_RV_A + V_A^2\right]}}$$

which expression reduces to

$$(2) \qquad \rho_{A\frac{R}{A}} = \frac{\rho_{AR}V_R - V_A}{\sqrt{V_R^2 - 2\rho_{RA}V_RV_A + V_A^2}}$$

Equation (9) thus gives the correlation between Speed and Accuracy for tests where Time is fixed. The Maximum value, or $\Pi_{A\frac{R}{A}} = +1.00$ is given for $\Pi_{AR} = +1.$. Thus if the pupils get every problem that they attempt right, Speed and Accuracy will be perfectly correlated. For zero correlations between Attempts and Rights, Speed and Accuracy are negatively correlated, the value approximating $-\frac{\sqrt{2}}{2}$. Finally, if the ratio $\frac{V_A}{V_R}$ is equal to the value of Π_{AR} , the correlation between Speed and Accuracy will be zero. Tables 5 and 19 indicate that these last relationships will hold very closely for intelligence test data. The ratios of the V_p' and the corresponding correlations are approximately .7

Formula (9) is very useful for obtaining the correlation between Speed and Accuracy from the single correlation table for Attempts and Rights. This table gives the values $\Pi_{RA}, \sigma_R, \sigma_A, M_R,$ and M_A which are all that are required to obtain $\Pi_{A\frac{R}{A}}$. The correlationswith Accuracy in Part I were computed from ratios $\frac{R}{A}$ obtained for each variate by division. About a third of the coefficients were then checked by the above formula. Substitution in equation (9) requires but a few moments, while the division for ratios alone takes about an hour for 50 cases. A great saving of time is , therefore, effected by the use of the above formula, especially if the constants in the formula are needed for other purposes.

If Attempts are fixed, Speed is measured by the recip-

rocal of the Time, and Accuracy by Rights. Equation (8) then, reduces to

$$\rho_{\frac{A}{T}\frac{R}{A}} = -\rho_{RT}$$

For reasoning test material of the True-False type, administered with Attempts constant, low negative correlations were obtained for Rights and Time, indicating that the correlation between Speed and Accuracy given by the last formula is positive and low. For 15 groups of about 20 pupils each, the average correlation $\rho_{\frac{A}{T}\frac{R}{A}} = .20 \pm .05$. For both types of test administration, then, Speed and Accuracy exhibit correlation that is zero or barely large enough to be significant.

b. The Validity of Simple Ratios as Scoring Indexes

It has just been shown that the ratios giving Speed and Accuracy are relatively independent measures of intelligence. The choice of the proper index will therefor depend upon criteria such as the purpose for which the measurements were made, the validity of the index, and its reliability. The question of validity will be taken up first.

The use of certain linear formulae has been justified on the basis of their validity or correlation with a criterion. This same principle may be applied to ratios. If a criterion, K, be substituted in Formula (C) in place of $\frac{Z}{W}$ an expression for the validity of $\frac{X}{Y}$ may be written in the form

$$(10) \qquad \rho_{K\frac{X}{Y}} = \frac{\rho_{KX}V_X - \rho_{KY}V_Y}{\sqrt{V_X^2 - 2\rho_{XY}V_XV_Y + V_Y^2}}$$

The correlation tables for r_{KX} and r_{KY} will furnish all of the data necessary for this formula and save the labor of calculating $r_{K\frac{X}{Y}}$ by the direct method of division. In general if $r_{K\frac{X}{Y}}$ is significantly higher than r_{KX} or r_{KY} its use in place of the simple variables is justified on the grounds of higher validity.

While score is not the best measure of validity for indexes which are so closely related to it, nevertheless the data in Table 5 will give a suggestion as to the general method. The order of the correlations between R, $\frac{R}{A}$, A, and score is $r_{SR}=.96$, $r_{S\frac{R}{A}}=.76$, $r_{SA}=.59$. The differences are all significant here so that on the basis of validity $\frac{R}{A}$ is a better index than A, but not so good as R alone.

c. The Reliability of Simple Ratios as Scoring Indexes

Equation (8) is again useful in predicting the reliability of a ratio without direct calculation on the ratios themselves. If $\frac{X_1}{Y_1}$ and $\frac{X_2}{Y_2}$ denote the ratios in question on successive trials of the same test or by parallel forms, the reliability formula may be written in the form

$$(11) \qquad r_{\frac{X_1}{Y_1}\frac{X_2}{Y_2}} = \frac{r_{X_1X_2}V_{X_1}V_{X_2} + r_{Y_1Y_2}V_{Y_1}V_{Y_2} - r_{X_1Y_2}V_{X_1}V_{Y_2} - r_{X_2Y_1}V_{X_2}V_{Y_1}}{\sqrt{\left[V_{X_1}^2 - 2r_{X_1Y_1}V_{X_1}V_{Y_1} + V_{Y_1}^2\right]\left[V_{X_2}^2 - 2r_{X_2Y_2}V_{X_2}V_{Y_2} + V_{Y_2}^2\right]}}$$

In order to calculate this quantity, four correlation tables are required: X_1X_2, X_1Y_2, X_1Y_1, and X_2Y_2.

If they are prepared all at once, the marginal frequencies give excellent checks on the distributions. As in the case of validity if $\Pi_{\frac{X_1}{Y_1}\frac{X_2}{Y_2}}$ is significantly greater than $\Pi_{X_1X_2}$ or $\Pi_{Y_1Y_2}$, it is to be preferred as the more reliable index. This method was applied to the Terman Scale Forms A and B. The results of the direct calculation appear in Table 7. Prediction by Formula (11) gives: $\Pi_{R_1R_2} = +.896$, $\Pi_{\frac{R_1}{A_1}\frac{R_2}{A_2}} = +.854$, and $\Pi_{A_1A_2} = +.684$ Accuracy and Rights are significantly more reliable than Attempts, but are not essentially different from one another.

The intelligence quotient, and similar ratios, are essentially a score divided by a chronological age, when the score is expressed in age units and taken from a suitable origin. The choice of the age unit is important chiefly because the resulting ratio is then a pure number and easily interpreted. As far as the validity and reliability of the ratio are concerned, the choice of the unit for score is of no consequence, inasmuch as correlation is a measure independent of the units employed for the two variates. The origin from which the score is taken is always of importance since any shift obviously changes the ratio e.g. $\frac{X}{A} \neq \frac{X+C}{A}$ The origin for the score in the intelligence quotient is the same as for chronological age, hence no difficulty is encountered.

Formula (8) is a function of Π's and V's, only the lat-

ter being affected by the origin from which the variables are taken. Indeed, by suitable choice of origin all of the $V's$ may be made equal, so that the formula reduces to

$$(12)\quad r_{\frac{X'}{Y'}\frac{Z'}{W'}} = \frac{r_{X'Z'} + r_{Y'W'} - r_{X'W'} - r_{Y'Z'}}{2\sqrt{(1 - r_{X'Y'})(1 - r_{Z'W'})}}$$

Letting Y' and W' denote age, and X' and Z' scores on successive trials of a test, an expression of the reliability of the ratio $\frac{S}{age}$ may be written in the form

$$r_{\frac{S_1}{age}\frac{S_2}{age}} = \frac{1 + r_{S_1S_2} - r_{S_1age} - r_{S_2age}}{2\sqrt{(1 - r_{S_1age})(1 - r_{S_2age})}}$$

Furthermore if $r_{S_1age} = r_{S_2age}$ this expression reduces to

$$(13)\quad r_{\frac{S_1}{age}\frac{S_2}{age}} = \frac{1 + r_{S_1S_2} - 2r_{Sage}}{2 - 2r_{Sage}}$$

Formula (13) will have the value +1.00 for $r_{SS} = +1.00$, and will be zero for $1 + r_{S_1S_2} = 2r_{Sage}$. For a given positive value of $r_{S_1S_2}$, the reliability given by the formula will increase as r_{Sage} decreases from the value $\frac{1 + r_{S_1S_2}}{2}$.

As an illustration, let $r_{S_1S_2} = +.8$. The function $r_{\frac{S_1}{age}\frac{S_2}{age}}$ may then be tabled for the argument r_{Sage} as follows:

$r_{\frac{S_1}{age}\frac{S_2}{age}}$	$r_{s\,age}$
0.	+.9
+.5	+.8
+.75	+.6
+.83	+.4
+.88	+.2
+.9	0.0
+.91	-.2
+.93	-.4
+.94	-.6
+.94	-.8
+.95	-.9

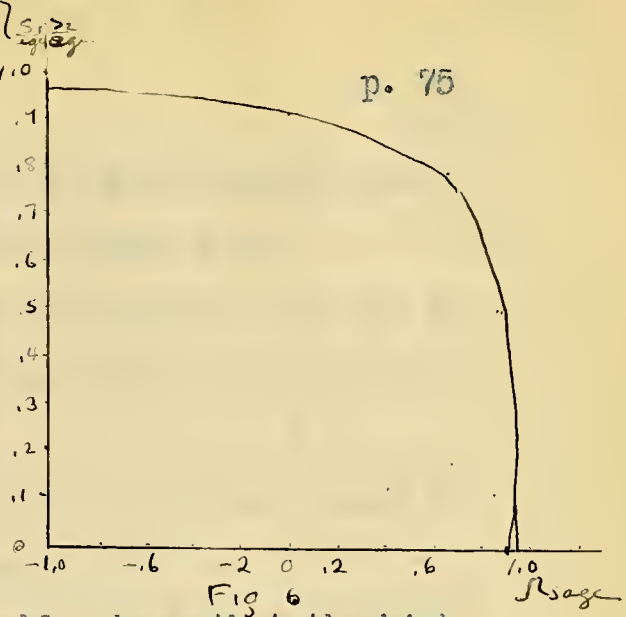


Fig 6

Figure 6 which is based on the table shows that the highest reliability of the ratio $\frac{S}{age}$ occurs with the highest negative values of $r_{s\,age}$ and decreases as $r_{s\,age}$ increases (to the right). The value for which $r_{\frac{S_1}{age}\frac{S_2}{age}} = .8$ is found by solving the equation

$$.8 = \frac{1.8 - 2x}{2 - 2x}$$

giving $x = .5$, so that the ratio has a greater reliability than $r_{S_1 S_2}$ up to the value $r_{s\,age} = .5$. Inasmuch as .8 is a good reliability coefficient for score, and $r_{s\,age}$ is seldom as high as +.5, the suppostitious example above indicates that ratios with age have in general a greater reliability than that between crude scores. The negative correlation usually found between score and age for a given grade group (See Tables 8 and 9) also indicates that ratios of the above type are most reliable when the group is thus selected with respect to age.

A final example will be given to illustrate the effect of transferring the origin to eliminate the . Assuming the approximate values $V_A = V_B = 20$, $r_{AB} = .9$, $V_{age} = 8$ and $r_{s\,age} = -.4$, formula (11) will give by simple substitution

$$r_{\frac{A}{age}\frac{B}{age}} = +.93$$

Shifting the origin to eliminate the V's and applying formula (13) gives

$$r_{\frac{A}{age}\frac{B}{age}} = +.96$$

An increase in reliability of .03 is thus brought about by the transfer of origin. For the linear form $S = a\,R + c\,W$ such a shift will, of course, have no effect upon the reliability or validity of the formula (See Theorem 1 Appendix).

The bases for selecting a simple ratio as a mode of indexing may be listed as follows:

1. The desirability of indexing another feature of the general characteristic, e.g. Accuracy, though a sub- characteristic of intelligence is essentially different from Speed.
2. The general properties of the ratio as compared with other variables e.g. the intelligence quotient facilitates comparisons with normal achievements.
3. The validity of the ratio to be determined by Formula (10).
4. The reliability of the ratio as determined by formula (11).

Section 20 General Conclusions

1. The various types of response to test material have been treated as index variables for the traits in question. An analysis of these variables for intelligence test data

revealed fairly definite relationships between them as indicated by the coefficient of correlation.

2. By eliminating the difficulty factor, the primary index variables were reduced to A,R,W, and T, one of these being fixed by the method of test administration.

3. An analysis of whole scales indicated that all of the primary variables have valuable properties as indexes. The introduction of the simple ratio $\frac{R}{A}$ made possible a comparison of the indexes revealing them in order of general reliability as S,R,$\frac{R}{A}$,W, and A, with Time fixed.

4. According to the criteria employed, the complicated formulae used by the authors of the tests, are not significantly better than Rights alone. Accuracy comes next as a generally reliable index, and has properties which the other variables do not possess. For batteries of tests, then, scoring by Rights alone is justified by reason of greater simplicity and practically equal reliability as compared with more complicated formulae.

5. In discriminative capacity, Attempts proved to be highest and Accuracy lowest; i.e. individuals and groups differ more widely in Speed than in Accuracy. Lack of discrimination between groups is of less consequence in an index than failure to differentiate between individuals. Accuracy, therefore, retains its high place as an

index regardless of the slight inter-group differences shown.

6. Analysis of the scales by component tests furnished a check upon the results obtained for whole batteries. The coefficients in general are lower and less consistent than by whole scales.

7. The validity and reliability of tests are increased by pooling components. Estimations of these correlations are furnished by certain predictive formulae, which in general, tend to give over estimation very early in the cumulated series.

8. Both theoretical prediction and actual results indicate that pooling tests soon ceases to increase validity and reliability materially. A battery of five well-selected tests is about as satisfactory as one twice as long.

9. The formula $S = aR + cW$ has been shown to be the most general linear form. Formulae for validity have been worked out by the method of least squares.

10. The linear formula above is open to ~~question~~ objection because of its sensitiveness for values of C. This implies that a new determination is necessary for each new group dealt with.

11. The scoring formula $R - W$ is criticized because of its

failure to correct for guessing. It has been shown that for Attempts constant (the usual method with True-False tests) corrective formulae of the type $R - \frac{1}{k}W$ have exactly the same validity as Rights alone.

12. Special formulae and methods have been worked out for determining the validity and reliability of simple ratios.

13. The valuable properties possessed by simple ratios indicate that they are highly desirable and useful scoring devices in spite of the labor of division. Special tables give such quotients directly.

14. More complicated formulae have not been dealt with because the labor involved in their use would be prohibitive no matter what virtues they might be found to possess. The sensitiveness of such formulae is additional reason for avoiding them.

15. The results as a whole point to the conclusion that for batteries, and for single tests as well, the most desirable and useful indexes are R and $\frac{R}{A}$.

Bibliography

Memoirs of the National Academy of Sciences, Vol. XV, Psychological Testing in the United States Army.

C.Spearman, Correlation of Sums and Differences, British Journal of Psychology, Vol. V, pp. 419-426.

A.S.Otis, An Absolute Point Scale for the Group Measurement of Intelligence. Journal of Educational Psychology, Vol. IX, Nos.5 and 6, May-June, 1918

L.L.Thurstone, A Scoring Method for Mental Tests, Psy. Bull., Vol.XVI, No. 7, July 1919.

T.L.Kelley, The Reliability of Test Scores Jl.Ed.Research Vol.III, No.5,p.379,May 1921

A.S.Otis and H.E.Knollin, The Reliability of the Binet Scale and Pedagogical Scales, Jl.Ed. Research., Vol. IV, No.2 121, Sept. 1921

Bibliography --Texts

G.U.Yule, An Introduction to the Theory of Statistics, Charles Griffin and Company, London, 1.1.v.

A.L.Bowley, Elements of Statistics, Charles Scribner's Sons, New York, 1907

D.C.Jones, A First Course in Statistics, G.Bell and Sons London, 1931

W.Brown and G.H.Thompson, The Essentials of Mental Measurement, Cambridge University Press, London, 1921

APPENDIX A

Correlation Tables for Reliability Coefficients

CORRELATION TABLE FOR SCORE ON TERMAN FORM A WITH SCORE ON TERMAN FORM B WITH GROUP I HIGH C

SCORE ON TERMAN FORM B (FIRST TRIAL)

SCORE ON TERMAN FORM A (SECOND TRIAL)	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100-109	110-119	120-129	130-139	140-149	150-159	160-169	170-179	180-189	f
200-209																	1	1
190-199																		
180-189																		
170-179																		
160-169														1	1			1
150-159												1	2	2	1			6
140-149									2	3			2	1				8
130-139								1	1	1	3	2						8
120-129				1			1	3	2	2		1						10
110-119					1	2	4	5	3	3								18
100-109						2	7	7	1									17
90-99			1		3	3	5	1										13
80-89			2	3	6	3	2											16
70-79		4	2	6	2	3	1											18
60-69		3	1	2	1	1												8
50-59		2	2															4
40-49	1	1	2															4
30-39	1																	1
20-29	1																	1
f	3	10	10	12	13	14	20	17	9	9	3	4	4	4	2	-	1	135

$R = +.908 \pm .010$

$M_A = 101.74 \pm 1.79$

$M_B = 84.11 \pm 1.96$

$M_A - M_B = 17.63 \pm 2.65$

$\therefore$ gain is significant

$\sigma_B = 33.79 \pm 1.39$

$\sigma_A = 30.91 \pm 1.27$

$\sigma_B - \sigma_A = 2.88 \pm 1.88$

diff. may be significant

CORRELATION TABLE FOR ATTEMPTS ON TERMAN FORM A WITH ATTEMPTS ON TERMAN FORM B WITH GROUP I HIGH C

ATTEMPTS ON TERMAN FORM B (FIRST TRIAL)

ATTEMPTS ON TERMAN FORM A	80–89	90–99	100–109	110–119	120–129	130–139	140–149	150–159	160–169	170–179	180–189	f
180–189			1					2	6	8	9	26
170–179					1	2	3	5	11	8		30
160–169				1	4	5	4	7	6	4		31
150–159	1			3	1	5	6	3	1			20
140–149			1	3	2	3	1	2				12
130–139	1			1	1	1		1				5
120–129			1	1	1		1					4
110–119		1	1	1	1	2						6
100–109								1				1
90–99		1										1
80–89	1											1
f	3	2	4	10	11	18	17	21	28	20	9	135

$R = +.684 \pm .032$

$M_A = 161.00 \pm 1.20$

$M_B = 148.19 \pm 1.40$

$M_A - M_B = 12.81 \pm 1.84$

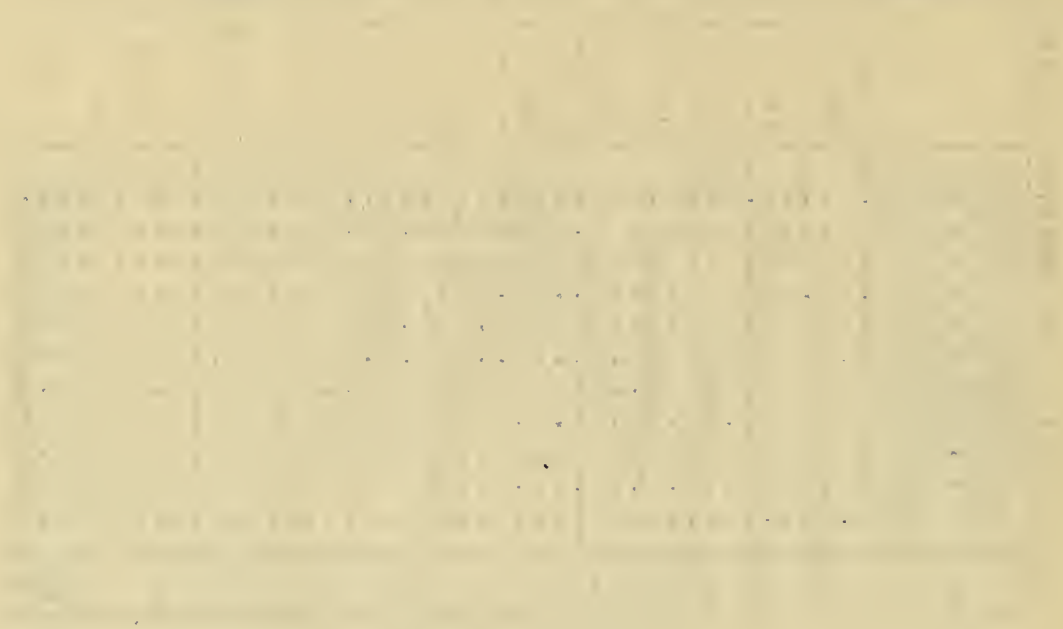
$\sigma_B = 24.12 \pm 0.99$

$\sigma_A = 20.63 \pm 0.85$

$\sigma_B - \sigma_A = 3.49 \pm 1.30$

∴ gain is significant

∴ diff. may be significant



CORRELATION TABLE FOR RIGHTS ON TERMAN FORM A WITH RIGHTS ON TERMAN FORM B WITH GROUP I HIGH C

RIGHTS ON TERMAN FORM B (FIRST TRIAL)

RIGHTS ON TERMAN FORM A	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100-109	110-119	120-129	130-139	140-149	150-159	160-169	f
170-179														1	1
160-169															-
150-159												1			1
140-149											3	2			5
130-139								1	3	2	3	1			10
120-129					1		3	6	3		1				14
110-119				1	1	2	6	4	4						18
100-109					2	8	11	4	1						26
90-99				6	3	6	2								17
80-89		2	6	3	4	4									19
70-79	1	1	6	4	2										14
60-69	1	2	3												6
50-59		2	1												3
40-49															-
30-39		1													1
f	2	6	16	14	13	20	22	15	11	2	7	4	-	1	135

$r = +.896 \pm .011$

$M_A = 102.41 \pm 1.37$

$M_B = 87.15 \pm 1.56$

$M_A - M_B = 15.26 \pm 2.08$

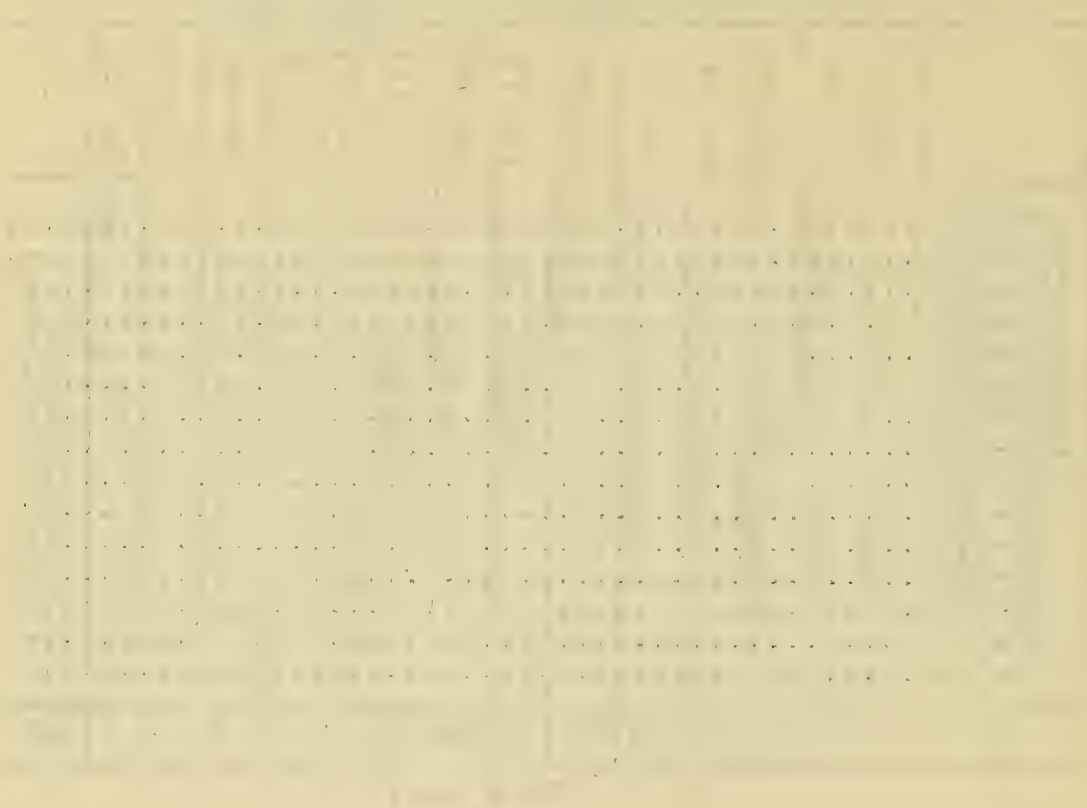
$\sigma_B = 26.95 \pm 1.11$

$\sigma_A = 23.54 \pm 0.97$

$\sigma_B - \sigma_A = 3.41 \pm 1.47$

∴ gain is significant

∴ diff. may be significant



CORRELATION TABLE FOR WRONGS ON TERMAN FORM A WITH WRONGS ON FORM B WITH GROUP I HIGH C

WRONGS ON TERMAN FORM B (FIRST TRIAL)

WRONGS ON TERMAN FORM A	10-19	20-29	30-34	40-49	50-59	60-69	70-79	80-89	90-99	100-109	110-119	f
100-109							1					1
90-99					1	2	3		1		1	7
80-89				2		3	5	3	3	2		16
70-79					1	7	9	3	1	1		22
60-69				3	3	6	3	3		1		18
50-59		1		7	12	5	4					29
40-49			4	8	4	2	3					21
30-39		1	5	3	2	1						11
20-29		4	4									8
10-19	1	1										2
f	1	7	13	23	22	25	28	9	4	4	1	135

$r = +.737 \pm .027$

$M_A = 60.00 \pm 1.14$

$M_B = 60.34 \pm 1.14$

$M_A - M_B = -0.34 \pm 1.60$

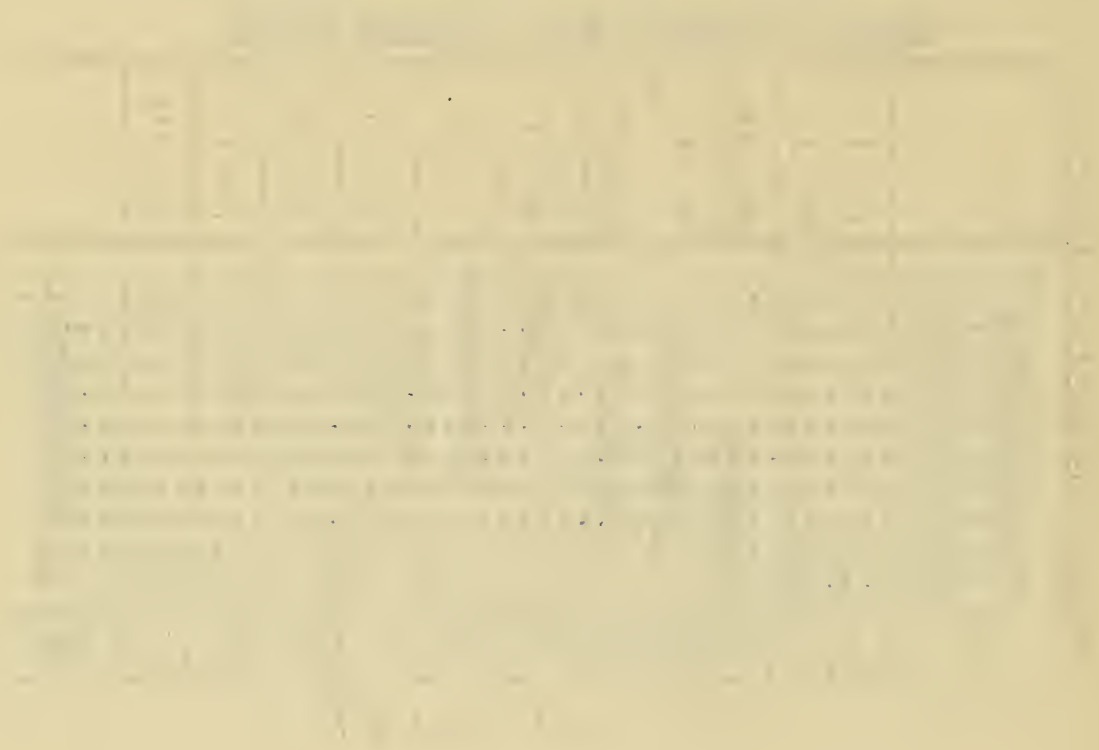
$\therefore$ loss is insignificant

$\sigma_B = 19.73 \pm 0.81$

$\sigma_A = 19.66 \pm 0.81$

$\sigma_B - \sigma_A = 0.07 \pm 1.15$

$\therefore$ diff. is insignificant



CORRELATION TABLE FOR ACCURACY ON TERMAN FORM A WITH ACCURACY ON FORM B WITH GROUP I HIGH C

ACCURACY ON TERMAN FORM B (FIRST TRIAL)

ACCURACY ON TERMAN FORM A	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	f
90-95														1	1
85-89												1			1
80-84								1		2	3	3	1		10
75-79							1	2	2	2	4	2			13
70-74							4	5	3	3					15
65-69					1		3	6	4	3	4				20
60-64						1	3	6	6	3	1				20
55-59					4	2	5	6	2	2					21
50-54				4	4	4	1	1	1						15
45-49				6	1	3		2							12
40-44				2	1		1								4
35-39			1	1											2
30-34	1														1
f	1	1	13	11	10	13	30	32	15	13	7	6	1	1	135

$r = +.841 \pm .017$

$M_A = 0.633 \pm 0.007$

$M_B = 0.585 \pm 0.008$

$M_A - M_B = 0.048 \pm 0.011$

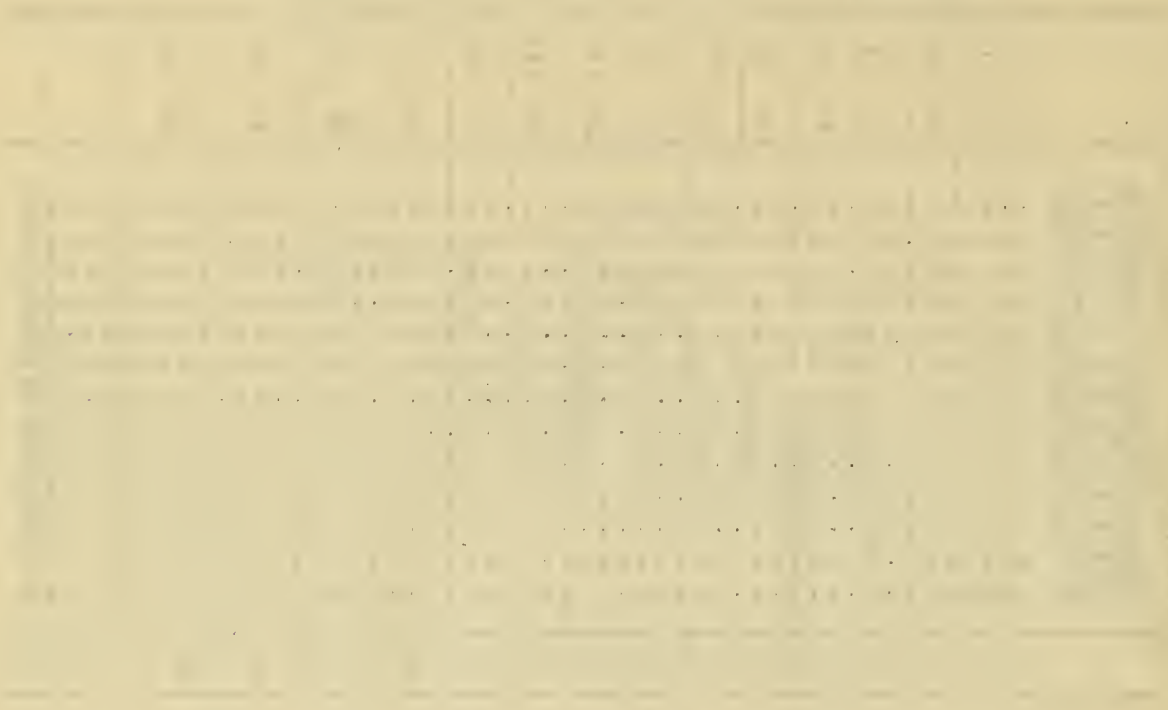
$\sigma_B = 0.1322 \pm 0.0054$

$\sigma_A = 0.1237 \pm 0.0051$

$\sigma_B - \sigma_A = 0.0085 \pm 0.007$

∴ gain is significant

∴ diff. is not significant



APPENDIX B

Theorems Relating to Correlation

Notations:

X, Y, Z -- variables from arbitrary origins

x, y, z --- variables from respective means

M_x, ---- means of the variables X, ---

σ_x, ---- standard deviations of X, ---.

V_x, ----- Pearson's coefficient of variability, $\frac{100\sigma}{M}$

ρ_{xy} ---- product moment correlation

Σ ----- sum of such quantities as ----.

N ---- frequency of the population

a, b, c, -- constants

Theorem 1. The correlation between two variables is the same as that between any two linear functions of each of them i.e.

(1) $\rho_{(aX+b)(cY+d)} = \rho_{xy} \quad (a, c \neq 0)$

Transferring the variables to their respective means,

$$\rho_{(aX+b)(cY+d)} = \rho_{(ax)(cy)}$$
$$= \frac{\Sigma axcy}{\sqrt{\Sigma a^2x^2 \Sigma c^2y^2}}$$
$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$
$$= \rho_{xy}$$

Theorem 2. The correlation between two ratios $\frac{X}{Y}$ and $\frac{Z}{W}$ is given by the formula,

$$(2)\ \rho_{\frac{X}{Y}\frac{Z}{W}} = \frac{\rho_{XZ}V_XV_Z + \rho_{YW}V_YV_W - \rho_{XW}V_XV_W - \rho_{YZ}V_YV_Z}{\sqrt{\left[V_X^2 - 2\rho_{XY}V_XV_Y + V_Y^2\right]\left[V_Z^2 - 2\rho_{ZW}V_ZV_W + V_W^2\right]}}$$

The means and standard deviations of $\frac{X}{Y}$ are given by

equations (9) and (10) in Yule,[a]

$$(a) \qquad M_{\frac{x}{y}} = \frac{M_x}{M_y}\left(1 - \rho_{xy} V_x V_y + V_y^2\right)$$

$$(b) \qquad \sigma^2_{\frac{x}{y}} = \frac{M_x^2}{M_y^2}\left(V_x^2 - 2\rho_{xy} V_x V_y + V_y^2\right)$$

The required correlation $\rho_{\frac{x}{y}\frac{z}{w}}$ will then be given by

$$N \rho_{\frac{x}{y}\frac{z}{w}} \sigma_{\frac{x}{y}} \sigma_{\frac{z}{w}} = \Sigma\left(\frac{x}{y} - M_{\frac{x}{y}}\right)\left(\frac{z}{w} - M_{\frac{z}{w}}\right)$$

$$= \Sigma\left(\frac{x}{y}\frac{z}{w}\right) - N M_{\frac{x}{y}} M_{\frac{z}{w}}$$

$$= \frac{M_x}{M_y}\frac{M_z}{M_w}\Sigma\left(1+\frac{x}{M_x}\right)\left(1+\frac{y}{M_y}\right)\left(1+\frac{z}{M_z}\right)^{-1}\left(1+\frac{w}{M_w}\right)^{-1} - N M_{\frac{x}{y}} M_{\frac{z}{w}}$$

Expanding, neglecting terms higher then the second degree and substituting form (a) and (b) gives (2).

Formula (2) gives satisfactory results for fairly long series, but for very short ones, considerable error occurs due no doubt to neglecting the higher powers in the expansions of the binomials.

Theorem 3. If $X = aY + bZ$, the partial correlations between the variables may be written,

$$(3) \qquad \begin{aligned} \rho_{xy.z} &= +1, \\ \rho_{xz.y} &= +1, \\ \rho_{yz.x} &= -1. \end{aligned}$$

as is evident by inspection.

Example: Since $A = R + W$, $\rho_{RW} = -1$, if $A = \text{Const.}$

---

a. G.U.Yule, Introduction to the Theory of Statistics, pp. 214-16.

Theorem 4. The correlation between a variable and the sum of n others is given by the formula[a],

$$(4)\quad \rho_{(x_1+x_2+\cdots x_n)y} = \frac{\rho_{x_1y}\sigma_{x_1} + \rho_{x_2y}\sigma_{x_2} + \cdots + \rho_{x_ny}\sigma_{x_n}}{\sqrt{\sigma_{x_1}^2+\sigma_{x_2}^2+\cdots+\sigma_{x_n}^2+2\left(\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2}+\cdots \frac{n(n-1)}{2}\text{ terms}\right)}}$$

(this expression is a special case of Spearman's General Formula)

$$\rho_{(x_1+x_2+\cdots x_n)y} = \frac{\Sigma x_1y + \Sigma x_2y + \cdots + \Sigma x_ny}{N\sigma_{(x_1+x_2+\cdots x_n)}\sigma_y}$$

Since $\sigma_{(x_1+x_2+\cdots x_n)} = \sqrt{\sigma_{x_1}^2+\sigma_{x_2}^2+\cdots\sigma_{x_n}^2+2\left(\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2}+\frac{n(n-1)}{2}\text{terms}\right)}$

and $\Sigma xy = N\rho_{xy}\sigma_x\sigma_y$, the right hand member above reduces to the expression in (4)

Corollary. If the standard deviations σ_x are all equal

$$\rho_{(x_1+x_2+\cdots x_n)y} = \frac{\rho_{x_1y}+\rho_{x_2y}+\cdots+\rho_{x_ny}}{\sqrt{n+2\left(\rho_{x_1x_2}+\rho_{x_1x_3}+\cdots\frac{n(n-1)}{2}\text{terms}\right)}}$$

Theorem 5. The correlation between the sum of n variables and n other variables is given by the formula[b]

$$(5)\quad \rho_{(x_1+x_2+\cdots x_n)(x_1'+x_2'+\cdots x_n')} = \frac{\rho_{x_1x_1'}\sigma_{x_1}\sigma_{x_1'}+\rho_{x_1x_2'}\sigma_{x_1}\sigma_{x_2'}+\cdots\text{ to }n^2\text{ terms}}{\sqrt{\left[\sigma_{x_1}^2+\cdots\sigma_{x_n}^2+2\left(\rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2}+\cdots\right)\right]\left[\sigma_{x_1'}^2+\sigma_{x_n'}^2+2(\cdots)\right]}}$$

The proof is similar to (4)

Corollary 1. If the standard deviations σ_x are all equal

$$(6)\quad \rho_{nn} = \frac{\rho_{x_1x_1'}+\rho_{x_1x_2'}+\cdots n^2\text{ terms}}{\sqrt{\left[n+2\left(\rho_{x_1x_2}+\cdots\frac{n(n-1)}{2}\text{ terms}\right)\right]\left[n+2\left(\rho_{x_1'x_2'}+\cdots\frac{n(n-1)}{2}\text{terms}\right)\right]}}$$

where ρ_{nn} denotes the left hand member of (5).

---

a. C.Spearman, British Journal of Psychology. 1913, Vol.V, p.417.

b. C.Spearman, loc. cit.

Corollary 2. If the correlations ρ_{xx} are all equal, equation (6) may be written

$$\rho_{nn} = \frac{n\rho_{xx}}{1+(n-1)\rho_{xx}} \tag{7}$$

This is Brown's Theorem, but as shown above it is merely a special case of Spearman's General Formula.